Asymptotic theory for differentiated products demand models with many markets

Joachim Freyberger
Department of Economics, University of Wisconsin, Madison, United States

Abstract

This paper develops asymptotic theory for differentiated product demand models with a large number of markets $T$. It takes into account that the predicted market shares are approximated by Monte Carlo integration with $R$ draws and that the observed market shares are approximated from a sample of $N$ consumers. The estimated parameters are $\sqrt{T}$ consistent and asymptotically normal as long as $R$ and $N$ grow fast enough relative to $T$. Both approximations yield additional bias and variance terms in the asymptotic expansion. I propose a bias corrected estimator and a variance adjustment that takes the leading terms into account. Monte Carlo simulations show that these adjustments should be used in applications to avoid severe undercoverage caused by the approximation errors.

1. Introduction

Discrete choice models have been widely used in the empirical industrial economics literature to estimate demand for differentiated products. In these models, consumers in market $t$ can typically choose one of $J_t$ products or an outside option. The consumers choose the product that maximize their utility, which leads to an expression of the markets shares. The parameters of the utility function can then be estimated using observed markets shares and product characteristics.

The challenge of the asymptotic theory in these models is to deal with several approximation errors. First, in models with heterogeneous consumers, such as the model of Berry et al. (1995) (referred to as the BLP model), the market shares involve integrals over the distribution of random coefficients. When estimating the parameters of the model, these integrals cannot be calculated analytically and are usually approximated by Monte Carlo integration with $R$ random draws from the distribution of the random coefficients. Second, many data sets do not contain true market shares but approximations from a sample of $N$ consumers. Consequently, asymptotic theory needs to allow for three sources of errors: the simulation error in approximating the shares predicted by the model, the sampling error in estimating the market shares, and the underlying model error.

The limiting distribution of the estimated parameters can be obtained by either letting the number of products, the number of markets, or both approach infinity. Since the asymptotic distribution of the estimator serves as an approximation of its unknown finite sample distribution, it depends on the particular data set which approximation is most suitable. While in some cases using an approximation where the number of products approaches infinity is appropriate (as in Berry et al., 1995), in other cases the number of markets is a lot larger than the number of products (e.g. Nevo, 2001). As shown in this paper the asymptotic properties of the estimator differ a lot depending on which approximation is

---

*Corresponding author. E-mail address: jfreyberger@ssc.wisc.edu.*

http://dx.doi.org/10.1016/j.jeconom.2014.10.009
© 2014 Elsevier B.V. All rights reserved.

---

1 Although the focus lies on the effect of using Monte Carlo integration to approximate integrals, polynomial-based quadrature rules are also discussed. See Remark 4 for further details.
used. It is therefore important that both approximations are well understood.

Berry et al. (2004) provide asymptotic theory for a large number of products in one market. In their paper all market shares go to 0 at the rate 1/J and a necessary condition for asymptotic normality is that \( J^2/R \) and \( J^2/N \) are bounded. In this case, they find that their estimator \( \hat{\theta} \) of the parameter vector \( \theta_0 \) satisfies

\[
\sqrt{J} \left( \hat{\theta} - \theta_0 \right) \xrightarrow{d} N \left( 0, V_{\text{GMM}} + \lambda_1 V_{\text{MC}} + \lambda_2 V_{\text{SM}} \right),
\]

where \( \lambda_1 = \lim_{J \to \infty} \sqrt{J}^2/R \) and \( \lambda_2 = \lim_{J \to \infty} \sqrt{J}^2/N \). Here \( V_{\text{GMM}} \) denotes the variance of the estimator when the integral is calculated exactly and the true market shares are observed. \( V_{\text{MC}} \) and \( V_{\text{SM}} \) are additional variance terms due to the approximation error of the integrals and the market shares, respectively. Hence, if \( J^2/R \) or \( J^2/N \) are bounded away from 0, the asymptotic distribution of the estimated parameter vector is centered at 0 but Monte Carlo integration and market share approximation leads to a larger variance.

This paper is concerned with the asymptotic theory for a small number of products, \( J \), in a growing number of markets \( T \), which is the natural choice in many applications. However, this setup has not been considered in the literature so far. In Nevo (2001), for example, there are 1124 markets and 24 products (see also among others Kim, 2004, and Villas-Boas, 2007). Furthermore, in a similar (but more general) class of models, Berry and Haile (2014) provide nonparametric identification results for a large number of markets and a fixed number of products. These results can serve as a basis for nonparametric or semiparametric estimation of the model. Before such a flexible estimation procedure is developed, it is interesting to know how the commonly used fully parametric estimators behave. I prove consistency and asymptotic normality for these cases where \( T \) approaches infinity and \( J \leq \tilde{J} \) where \( \tilde{J} \) is fixed. The assumptions in the main part of the paper are stated for the BLP model, but the results in the appendix use higher level assumptions for a more general class of models. I find that the estimated parameters are \( \sqrt{T} \) consistent and asymptotically normal as long as \( \sqrt{T}/R \) and \( \sqrt{T}/N \) are bounded. In this case, \( \hat{\theta} \) satisfies

\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) \xrightarrow{d} N \left( \tilde{\lambda}_1 \mu_1 + \tilde{\lambda}_2 \mu_2, V_{\text{GMM}} \right)
\]

where \( \tilde{\lambda}_1 = \lim_{T \to \infty} \sqrt{T}/R \) and \( \tilde{\lambda}_2 = \lim_{T \to \infty} \sqrt{T}/N \), again \( V_{\text{GMM}} \) is the variance of the estimator without approximation errors in the integrals and market shares, and \( \mu_1 \) and \( \mu_2 \) are constants. Hence, if \( \sqrt{T}/R \) or \( \sqrt{T}/N \) are bounded away from 0, Monte Carlo integration and market share approximation lead to an asymptotic distribution which is not centered at 0. The intuition for this result is that Monte Carlo integration (and similarly market share approximation) yields both additional bias as well as additional variance terms in the asymptotic expansion. The leading bias term is of order \( O_p(\sqrt{T}/R) \), while the leading variance term is of order \( O_p(1/\sqrt{T}) \). Hence, the bias dominates the variance if \( R \) does not increase faster than \( T \), and the variance does not affect the asymptotic distribution as long as \( R \to \infty \). These results rely on using different draws to approximate the integral in different markets. If the same \( R \) draws are used in all markets one needs \( T/R \) to be bounded to obtain \( \sqrt{T} \) consistency, which means that more draws are needed to approximate each integral relative to the number of markets.²

These results highlight that there are important differences between letting the number of products or the number of markets approach infinity. With a large number of products, it is important to correct the variance of the estimator due to the approximations. With a large number of markets and different draws in each market, the asymptotic distribution might not be centered at 0 and hence, a bias corrected estimator is needed. In both cases if \( R \) is too small, confidence intervals based on the usual asymptotic GMM distribution have the wrong size even asymptotically. Notice that when \( \sqrt{T}/R \) and \( \sqrt{T}/N \) converge to 0, the approximations do not affect the asymptotic distribution of the estimator. Contrary, in the setup where \( f \) approaches infinity, this result is only obtained when \( J^2/R \) and \( J^2/N \) converge to 0, which is a stronger requirement on the number of draws relative to the sample size.

The finite sample properties of the estimator depend on \( R \) and \( N \) due to both the additional bias and variance terms. I suggest an analytical bias correction which eliminates the leading bias term from the asymptotic distribution. I also show how one can easily incorporate the leading variance term when calculating standard errors. These two corrections greatly improve finite sample results. In particular, Monte Carlo simulations demonstrate that using a small number of simulation draws in comparison to the number of markets and using the usual GMM asymptotic distribution can yield distorted inference while the use of bias correction and adjusted standard errors leads to a considerably better performance.

These results might suggest that practitioners can simply use a very large number of draws and ignore Monte Carlo integration issues. Although this might be feasible depending on the model and computing resources available, in applications this is often not possible for several reasons. First, one does not know in advance how many draws suffice to obtain satisfactory results. As discussed in Section 4, the number of draws needed depends, among others, on the sample size, the number of random coefficients as well as unknown parameters, such as the variance of the random coefficients. Second, taking a very large number of draws is computationally very demanding because one needs to solve a complicated nonlinear optimization problem to estimate the parameters. The Monte Carlo results of the random coefficients logit model presented in Section 4 are based on a small number of products (\( J = 4 \) and six random coefficients to make the problem tractable. However, in the same setup as in Section 4 but with a sample size of \( J = 24 \) and \( T = 1124 \) (as in Nevo, 2001) it takes more than 24 h to minimize the objective function when \( R = 2000 \) and the starting values of the parameters are close to the true values. Since we are dealing with a nonlinear optimization problem one needs to use several different starting values in applications. Taking a smaller number of draws considerably speeds up calculations. Third, even when taking the same draws for each product and each random coefficient, the number of draws needed is \( T \times R \). In the previous example this means that 2248,000 draws are used to calculate the shares and the draws have to be stored before optimizing the function. As a consequence, more than 20 GB of memory is needed to run the program which is used to do the simulations in this paper. Finally, in case one wants to integrate over empirical distributions of demographic characteristics, \( R \) is constrained by the number of people in the database for a certain market.

The implication for empirical work with Monte Carlo integration or approximated market shares is that bias corrections and variance adjustments should be used. If the number of draws and the number of consumers is sufficiently large, the bias correction is close to 0 and the corrected standard errors will be very close to the

² Remark 3 in Section 2.2 shows the differences in the asymptotic expansions when the same draws and different draws are used in each market. This section also highlights the differences between asymptotics in the number of markets and the number of products with and without different draws in different markets. See Remark 5 for details.

³ Computational details, including a description of the processors used, are presented in Section 4.
usual GMM standard errors. If the number of draws or consumers is small, the approximations will affect the finite sample performance of the estimator, the usual GMM standard errors underestimate the true variance, and the estimates are biased. In this case the proposed corrections, which can be computed easily, considerably improve the finite sample performance. Nevertheless, the number of draws and consumers should be as large as possible, subject to computational constraints and data availability, in order to improve the precision of the initial estimator, which is used to calculate the bias correction.

The focus of this paper is on understanding the asymptotic theory when Monte Carlo integration is employed because this is the method which has almost exclusively been used in applications. Furthermore, an advantage of Monte Carlo integration is that it can easily be used to integrate over distributions with an unknown functional form such as demographic characteristics as in Nevo (2001). An interesting alternative is to use polynomial-based quadrature rules recently advocated by Skranda and Judd (2011). These approximations are shown to perform well in simulations when integrating over a normal distribution. However, it is not clear how a distribution of demographic characteristics, which does not have a closed form expression, can be handled with quadrature rules unless a parametric distribution is assumed. Although the focus lies on Monte Carlo integration, the asymptotic expansions derived in this paper also provide insights into finite sample bias from polynomial-based quadrature rules.

The results in this paper are related to similar findings of Lee (1995) in simulated maximum likelihood estimation (SMLE) of discrete choice models. Lee’s paper and my paper are both based on higher order expansions of the derivative of the objective function in which the simulation error of the integral is linearized. The first term in this expansion is the derivative without the simulation error, the second term leads to a second variance term, and the third term to the bias. Lee’s paper is tailored to SMLE while mine is tailored to discrete choice random coefficient models. Lee (1995) also finds that the asymptotic distribution might not be centered at 0 if the number of draws is small relative to the sample size.4 Similar findings are obtained in nonlinear simulated least squares estimation by Laffont et al. (1995), where again the assumptions and proofs are specific to that setup. Simulated methods of moments estimators studied by McFadden (1989) and Pakes and Pollard (1989) differ from the previously discussed estimators because they do not lead to a bias term in the asymptotic expansion. Hahn and Newey (2004) and Arellano and Hahn (2007) propose bias corrections in nonlinear fixed effects panel data models where the bias is not due to a simulation error, but the incidental parameters problem and there is no variance adjustment.

Kristensen and Salanié (2013) deal with simulation based estimators in a general setup where the function to be approximated is the same for all \( t \). They assume that an expansion of the estimator exists and that the remainder term is negligible. Under these assumptions, they propose bias corrections which eliminate the leading bias term and variance adjustments. These results do not directly apply to the BLP model because both the distribution of the random coefficients and the true market shares might vary over markets. However, in case the true market shares are observed and the distribution of random coefficients is the same in all markets, the expansions and the bias correction here are a special case of those in Kristensen and Salanié (2013). In this case, I still need to derive the actual expansion, which is needed to calculate the bias correction and the variance adjustment, and show that the remainder term is negligible under low level conditions. Doing so is particularly challenging in discrete choice demand models because there is no closed form expression for the objective function and the resulting formulas are algebraically tedious. See Remark 2 for more details.

My results provide theoretical justifications for simulation results in recent studies by Skranda and Judd (2011) and Skranda (2012), who find among others finite sample bias and excessively tight standard errors when using Monte Carlo integration. I show in Monte Carlo simulations that my adjusted standard errors are not artificially tight and that the finite sample bias caused by the integral approximation can be eliminated by using the analytical bias correction. Other recent contributions to the literature on estimation of discrete choice demand models include Gandhi et al. (2013) and Armstrong (2013). In both papers \( J \to \infty \). Gandhi et al. (2013) allow for interactions between the unobserved demand error and product characteristics, which affect both the identification arguments and estimation method. Armstrong (2013) discusses the validity of commonly used instruments in models with a large number of products under conditions on economic primitives. He shows that in several models using product characteristics as instruments for price yields inconsistent estimates and he shows how consistent estimates can be obtained. His findings on invalidity of instruments do not apply in my setting with a fixed number of products.

The remainder of this paper is organized as follows. The next section introduces the BLP model, provides low level assumptions, and states the main results. A more general class of models under higher level assumptions is discussed in the Appendix. Section 3 explains how the bias correction and the variance adjustment can be implemented. Finally, I demonstrate with Monte Carlo simulations that the proposed corrections considerably improve the finite sample performance.

2. Random coefficient logit model

Let \( p_{jt} \in \mathbb{R} \) be the price of product \( j \) in market \( t \). Define \( x_{jt} = \left( p_{jt}, \xi_{jt}^{(2)} \right) \in \mathbb{R}^{d_\xi} \), where \( \xi_{jt}^{(2)} \) are observed product characteristics other than price and may include a constant or product dummies. Let \( \xi_{jt} \in \mathbb{R} \) be a product characteristic of product \( j \) in market \( t \), which is observed by firms and consumers, but not by the econometrician. In the BLP model consumer \( i \) in market \( t \) chooses the product \( j \) which maximizes the utility

\[
        u_{it} = x_{jt} \beta_{it} + \xi_{jt} + \epsilon_{jt},
\]

where \( \beta_{it} = \beta_0 + \Sigma_0 \theta_i \), \( \beta_0 \in \mathbb{R}^{d_\beta} \), \( \Sigma_0 \) is a \( d_\beta \times d_\beta \) matrix, and \( \theta_i \in \mathbb{R}^{d_\theta} \) is a random vector with distribution function \( P_{\theta_\ell} \). Assuming that \( \epsilon_{it} \) are independent and identically distributed (iid) extreme value random variables, the market share of product \( j \) in market \( t \) is given by

\[
        s_{jt} = \int \frac{\exp(x_{jt} \beta_0 + \xi_{jt} + \xi_{jt} \Sigma_0 v)}{1 + \sum_{k=1}^{J_t} \exp(x_{kt} \beta_0 + \xi_{kt} + \xi_{kt} \Sigma_0 v)} dP_{\theta_\ell}(v).
\]

I assume that \( P_{\theta_\ell} \) is known to the econometrician. In many applications \( P_{\theta_\ell} \) is assumed to be a standard normal distribution and \( \Sigma_0 \) is a diagonal matrix, but \( P_{\theta_\ell} \) could also be a distribution of demographic characteristics which varies across markets. Let \( x_i = (x_{i1}, \ldots, x_{iK}), \xi_i = (\xi_{i1}, \ldots, \xi_{iK}), \theta = (\beta, \text{vec}(\Sigma)) \), and define

\[
        v_{jt} (\theta, x_i, \xi_i, v) = \frac{\exp(x_{jt} \beta_0 + \xi_{jt} + \xi_{jt} \Sigma_0 v)}{1 + \sum_{k=1}^{J_t} \exp(x_{kt} \beta_0 + \xi_{kt} + \xi_{kt} \Sigma_0 v)}.
\]
The function \( v_{jt}(\theta, x_t, \xi_t, v) \) represents the markets share of product \( j \), predicted by the model, for a given parameters \( \theta \), product characteristics \( x_t \) and \( \xi_t \), and consumer type \( v \). Since for any market \( t \) the market shares are generated by integrating over all consumers, the shares predicted by the model are defined as

\[
\sigma_{jt}(\theta, x_t, \xi_t, P_t) = \int v_{jt}(\theta, x_t, \xi_t, v) \, dP_t(v)
\]

and the true shares are \( s_{jt} = \sigma_{jt}(\theta_0, x_t, \xi_t, P_{0t}) \). Let \( \sigma_{jt}(\theta, x_t, \xi_t, P_t) \) be the \( J \times 1 \) vector with elements \( \sigma_{jt}(\theta, x_t, \xi_t, P_t) \) and define \( v_{jt}(\theta, x_t, \xi_t, v) \) analogously.

Berry (1994) shows that for any \( \theta \in \Theta, x_t \in \mathbb{R}^{d_x \times k}, \) distribution \( P \), and \( s_{jt} \in (0, 1) \) there exists \( \xi_t(\theta, P, s_t, x_t) \in \mathbb{R}^J \) such that for all \( j \)

\[
s_{jt} = \int v_{jt}(\theta, x_t, \xi_t(\theta, P, s_t, x_t), v) \, dP_t(v).
\]

In particular, one can solve for \( \xi_t \) in the system of \( J \) equations given in (2) or, in other words, \( \sigma_t^{-1}(\theta, x_t, \xi_t, P_t) \) is invertible. I denote the inverse function by \( \sigma_t^{-1}(\theta, x_t, \xi_t) \).

Now assume that there are instruments \( z_t \in \mathbb{R}^{d_z \times N} \) such that

\[
E(\{z_t^\prime \xi_t(\theta_0, P_{0t}, s_t, x_t)\}) = 0
\]

and that these moment conditions identify \( \theta_0 \). Then an estimator of \( \theta_0 \) is

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} \left\{ \frac{1}{T} \sum_{t=1}^{T} z_t^\prime \xi_t(\theta, P_{0t}, s_t, x_t) \right\} W_T
\]

\[
\times \left( \frac{1}{T} \sum_{t=1}^{T} z_t^\prime \xi_t(\theta, P_{0t}, s_t, x_t) \right)^\prime,
\]

where \( W_t \) is a weighting matrix which converges in probability to a positive definite matrix \( W \). The estimator \( \hat{\theta} \) is infeasible in practice for two reasons. First, it is infeasible to calculate the integral in (2) analytically. Therefore, it is often approximated by \( R \) random draws from \( P_{0t} \). In particular,

\[
\int v_{jt}(\theta, x_t, \xi_t, v) \, dP_{0t}(v) \approx \frac{1}{R} \sum_{r=1}^{R} v_{jt}(\theta, x_t, \xi_t, v_{rt})
\]

\[
= \frac{1}{R} \sum_{r=1}^{R} \exp(x_{jt}^\prime \beta + \xi_t + x_{jt}^\prime \Sigma v) \, dP_{0t}(v),
\]

where \( v_{rt} \) are draws from the distribution \( P_{0t} \), which may differ across markets, and \( P_{0t} \) denotes the empirical distribution function. Second, often the true market shares \( s_t \) are not observed but are instead approximated from a sample of \( N \) consumers. That is, the observed market shares are \( s_{jm} = \frac{1}{N} \sum_{n=1}^{N} s_{jm} \), where \( s_{jm} \in (0, 1) \), and \( E(s_{jm}) = \hat{s}_{jt} \).

Now define the estimator used in applications

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} \left\{ \frac{1}{T} \sum_{t=1}^{T} z_t^\prime \xi_t(\theta, P_{0t}, s_t, x_t) \right\} W_T
\]

\[
\times \left( \frac{1}{T} \sum_{t=1}^{T} z_t^\prime \xi_t(\theta, P_{0t}, s_t, x_t) \right)^\prime.
\]

2.1. Assumptions

1. I now provide primitive sufficient conditions for the asymptotic normality results discussed in the introduction to hold in the BLP model. A more general class of models under higher level assumptions is discussed in the Appendix. Let \( J \) be a positive integer.

Assumption RC1. The data \( \{z_t, x_t, s_t\}_{t=1}^{T} \) are independent across markets and \( J \leq J \).

Assumption RC2. \( v_{jt} \) are iid draws over \( r \) and independent over \( t \) from \( P_{0t} \) and independent of \( (x_t, z_t, s_t) \), \( \ln(T)/R \rightarrow 0 \), and \( \ln(N)/N \rightarrow 0 \).

Assumption RC3. Let \( \bar{\theta} = \{\theta^1, \ldots, \theta^K\} \) be a finite set of distributions, \( v_t = g(u_t, w_t) \) where \( g : \mathbb{A} \times \mathbb{R}^{d_u} \rightarrow \mathbb{R}^{d_x} \) is continuous in both arguments, \( a_t \) is a vector of constants and \( \mathbb{A} \subset \mathbb{R}^K \) is a compact set, \( w_t \) is a random vector and the distribution of each element of \( w_t \) either \( (a) \) is in \( \bar{\theta} \) and is the same for all \( t \) or \( (b) \) its support is a subset of \( [M_l, M_u] \) for some constants \( M_l \) and \( M_u \).

Assumption RC4. \( s_{jm} \in (0, 1) \) are iid draws over \( n \) with \( E(s_{jm}) = s_t \), independent over \( t \), independent of \( (x_t, z_t, s_t) \), \( s_t \), and \( \ln(T)/N \rightarrow 0 \).

Assumption RC5. There exists \( \epsilon > 0 \) such that \( \epsilon \leq s_{jt} \leq 1 - \epsilon \) for all \( t = 1, \ldots, T \) and \( j = 0, \ldots, J \) with \( s_{0t} = 1 - \sum_{j=1}^{J} \hat{s}_{jt} \).

Assumption RC6. \( x_t \) is in a compact set. The parameter space \( \Theta \) is compact and \( \theta_0 \) is in the interior of \( \Theta \).

Assumption RC7. The instruments matrix has full rank, \( E \left( x_t^\prime \xi(\theta_0, P_{0t}, s_t, x_t) \right) \leq M \) for some \( M \) and \( E \left( x_t^\prime \xi(\theta_0, P_{0t}, s_t, x_t) \right) < \infty \) for all \( t = 1, \ldots, d_t \) and \( j = 1, \ldots, J_t \).

Assumption RC8. For all functions \( h_j(z_t, x_t, s_t, P_{0t}, J_t) \) with \( E(h_j(z_t, x_t, s_t, P_{0t}, J_t)) \leq M \),

\[
\lim_{t \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J_t} E \left( h_j(z_t, x_t, s_t, P_{0t}, J_t) \right)
\]

exists.

Assumption RC9. The parameter vector \( \theta_0 \) is identified from the moment conditions in (4). In particular, for all \( \delta > 0 \), there exists \( C(\delta) > 0 \) such that

\[
\lim_{t \rightarrow \infty} \text{Pr} \left( \inf_{\theta \neq \theta_0} \left\{ \frac{1}{T} \sum_{t=1}^{T} z_t^\prime \xi(\theta, P_{0t}, s_t, x_t) \right\} - \xi_t(\theta_0, P_{0t}, s_t, x_t) \right\} \geq C(\delta) = 1,
\]

where \( \| \cdot \| \) denoted the Euclidean norm, and the matrix \( \Gamma = \lim_{t \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} \partial \xi_t(\theta, P_{0t}, s_t, x_t) \)

has full rank.

Assumption RC1 states that the observed data is independent across markets and that the number of products in each market is bounded by \( J \). One way to allow for dependence across markets is to assume that \( \xi_t = \xi_t + \Delta \xi_t \) and to estimate the model with product dummies as in Nevo (2001). In this way product dummies can capture the part of the product characteristics that do not vary across markets. Assumption RC2 assumes that the draws used to
approximate the integral in each market are iid and independent of the observed data and that the number of draws diverge.\(^6\) Notice that I assume that different draws are used in different markets. This assumption does not lead to much larger computational costs compared to using the same draws in all markets, but it leads to a smaller variance of the estimator as shown in Remark 3. Moreover, the distribution of random coefficients needs to have four bounded moments. Assumption RC3 places additional restrictions on the distribution of random coefficients. If the distribution has compact support, this assumption does not restrict how it can vary across markets, which allows using demographics with compact support. It also allows that the distribution has unbounded support but in this case the way it varies over \(t\) is controlled by the finite dimensional parameter \(a_t\). For example, one could model demographics with a normal or a log normal distribution where the mean and the variance change across markets. Clearly, also the often used normal distribution with an unknown mean and variance is permitted.

Assumption RC4 says that the consumers are sampled randomly in each market and across markets and that the number of consumers grows with \(T\). Assumptions RC5 and RC6 state that the market shares are bounded away from 0 and 1 (including the shares of the outside option \(s_0\)) and that the product characteristics and parameters are in a compact set, respectively. This assumption is common in the simulation based estimation literature (for example McFadden (1989), Lee (1995), and Berry et al. (2004)). Assumption RC7 puts moment conditions on the instrument. Assumption RC8 is a limit condition which is needed because the data are not assumed to be identically distributed across market (in which case the assumption would follow immediately). The assumption holds for example if we assume that \(J_t\) and \(P_{0t}\) are random and iid across markets and that all other assumptions hold conditional on \(J_t\) and \(P_{0t}\). The assumption can also hold for deterministic \(J_t\) and \(P_{0t}\). It mainly rules out systematic changes of \(J_t\) or \(P_{0t}\) over \(t\). Finally, Assumption RC9 contains standard global and local identification conditions. Nonparametric identification with many markets and a finite number of products is shown by Berry and Halie (2014).

2.2. Asymptotic properties of the estimator

The previous assumptions lead to the following theorem which characterizes the asymptotic expansion of the estimator.

**Theorem 1.** Assume that Assumptions RC1–RC9 hold. Then there exist \(d_1 \times d_2\) matrices \(\Phi_1, \Phi_2, \Phi_3,\) as well \(d_1 \times 1\) vectors \(\mu_1\) and \(\mu_2\) such that

\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) = \left( (\Gamma'W\Gamma)^{-1} \Gamma'W + o_p(1) \right) \times \left( Q_{1T} + \frac{1}{\sqrt{R}} Q_{2T} + \frac{1}{\sqrt{N}} Q_{3T} + \frac{\sqrt{T}}{R} C_{1T} \right) \\
+ \frac{\sqrt{T}}{N} C_{2T} + o_p\left( \frac{\sqrt{T}}{R} \right) + o_p\left( \frac{\sqrt{T}}{N} \right)
\]

where

\[
Q_{1T} \xrightarrow{d} N(0, \Phi_1), \quad Q_{2T} \xrightarrow{d} N(0, \Phi_2), \quad Q_{3T} \xrightarrow{d} N(0, \Phi_3), \quad \text{and}
\]

\[
C_{1T} \xrightarrow{p} \mu_1, \quad C_{2T} \xrightarrow{p} \mu_2.
\]

Furthermore, \(Q_{1T}, Q_{2T},\) and \(Q_{3T}\) are asymptotically independent.

The proof is in the Appendix. It verifies higher level assumptions of a general class of models. An immediate consequence of Theorem 1 is the following result.

**Corollary 1.** Assume that Assumptions RC1–RC9 hold. If \(\lambda_1 = \lim_{T \to \infty} \frac{\sqrt{T}}{R} < \infty\) and \(\lambda_2 = \lim_{T \to \infty} \frac{\sqrt{T}}{N} < \infty\), then

\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) \xrightarrow{d} N \left( \left( (\Gamma'W\Gamma)^{-1} \Gamma'W (\lambda_1 \mu_1 + \lambda_2 \mu_2) \right), V_1 \right)
\]

where

\[
V_1 = (\Gamma'W\Gamma)^{-1} \Gamma'W \Phi_1 (\Gamma'W\Gamma)^{-1}.
\]

Theorem 1 shows that the use of Monte Carlo integration (as opposed to evaluating the integral exactly) leads to additional variance and additional bias terms. The leading variance term is of order \(1/\sqrt{R}\) while the leading bias term is of order \(\sqrt{T}/R\). Hence, if \(R\) grows more slowly than \(T\), the leading bias term dominates the leading variance term, which may lead to an asymptotic distribution that is not centered at 0. As a consequence, if \(\lambda_1 > 0\), confidence intervals based on the usual GMM asymptotic distribution have the wrong size asymptotically. If \(R\) grows faster than \(T\), the leading variance term becomes dominating, but the first order asymptotic distribution is not affected by Monte Carlo integration. Similar arguments apply to the approximations of the market shares.

**Remark 1.** The proof of Theorem 1 is based on a third order expansion of the derivative of the objective function in which the simulation error of the integral is linearized. This is analogous to Lee (1995), who makes use of a similar approach and obtains similar results for simulated maximum likelihood estimation. I also make use of similar tools as Lee to show that the remainder term of the expansion is negligible.\(^7\) Despite sharing these similarities at a general level, deriving the actual expansion and showing that the remainder term is negligible is significantly different in both models. In discrete choice demand models these results are challenging to derive (and very particular to this setting) because there is no closed form expression for the objective function. Moreover, the Taylor expansion is algebraically tedious because it involves differentiating the inverse of a function from \(R^k\) to \(R^l\). In the next subsection, I illustrate these steps for the special case where \(J_t = 1\).

**Remark 2.** Kristensen and Salanié (2013) deal with a general class of simulation based estimators, they use very similar expansions, and they obtain similar conclusions in their Theorem 2. However, Kristensen and Salanié (2013) do not allow the objects to be approximated to change over markets (here \(s_t\) and \(P_{0t}\)) and they directly assume that a higher order expansion exists and that the remainder term is negligible (in their Assumption A4). I derive these results, which depend on the specific structure of the model. As a consequence, I obtain closed form expressions for the variance and the bias, which are the basis for the adjustments proposed in Section 2.4. If the true market shares are observed and if the distribution of random coefficients is the same in all markets, the expansions and the bias correction here are a special case of those in Kristensen and Salanié (2013). In this case I still have to derive the actual expansion and show that the remainder term is negligible, which is the major part of the proof. The latter could be achieved by verifying Assumption A4 in Kristensen and Salanié (2013) using my low level conditions, but this condition is stronger than necessary. I instead show that the remainder term is \(o_p\left( \frac{\sqrt{T}}{R} + \frac{\sqrt{T}}{N} \right)\), which suffices to prove my results.

\(^6\) I assume throughout that the draws are random although in practice pseudo random draws are used. The implicit assumption is that pseudo random draws lead to the same asymptotic properties. The Monte Carlo simulation results support this assumption.

\(^7\) In particular, Lemmas A.2 and A.3 in the Appendix are very similar to Lemma A.2 and Proposition A.3 of Lee (1995), respectively.
Remark 3. It can also be shown that, under slightly different assumptions, if the distribution of random coefficients is the same in all markets and if one uses the same draws from $P_{Ok}$ in all markets then

$$\sqrt{T} \left( \hat{\theta} - \theta_0 \right) = \left( (\Gamma'W\Gamma)^{-1} \Gamma'W + o_p(1) \right) \times \left( Q_{1T} + \frac{\sqrt{T}}{\sqrt{R}} Q_{2T} + \frac{1}{\sqrt{N}} Q_{3T} + \frac{\sqrt{T}}{\sqrt{R}} C_{1T} \right) \left( \sqrt{T} \right) + o_p \left( \sqrt{\frac{T}{N}} \right),$$

where $Q_{2T}$ converges to a normally distributed random variable with mean 0 as well. In this case one needs $T/R$ to be bounded to obtain asymptotic normality, which is a stronger condition than the rate in Corollary 2, and the additional variance term dominates the additional bias term. Therefore, if $\hat{\lambda} = \lim_{T \to \infty} \sqrt{T} \left( \hat{\theta} - \theta_0 \right) \overset{d}{\to} N(0, V_1 + \hat{\lambda}^2 V_2)$

where $V_1 = \left( (\Gamma'W\Gamma)^{-1} \Gamma'W\phi, W\Gamma(\Gamma'W\Gamma)^{-1} \right)$ and $V_2 = \left( (\Gamma'W'\phi, W\Gamma(\Gamma'W'\phi)^{-1} \right)$. 

These results imply that using different draws in different markets is superior because the additional variance term is of smaller order while the bias is of the same order. This advantage of using different draws has been pointed out in other papers that deal with simulation based estimators, such as McFadden (1989) or Kristensen and Salanié (2013).

Remark 4. The expansions are very similar if one uses polynomial-based quadrature rules. However, the nodes are not random in these cases and, hence, one cannot use laws of large numbers or central limit theorems to deal with the terms $Q_{2T}$ and $Q_{3T}$. Since none of these terms has a mean of 0 without stochastic approximations, quadrature rules lead to an additional bias term only and one obtains asymptotic normality if the number of nodes grows fast enough relative to $T$. Quantifying the order of the bias is beyond the scope of this paper.

Remark 5. The main difference between the asymptotics in the number products (Berry et al., 2004) and the number of markets is that with many products all market shares are assumed to converge to 0, which leads to particular challenges for the asymptotic theory. Berry et al. (2004) assume that the same draws are used to approximate the predicted shares for all products and markets. Under this assumption, with many products $\frac{1}{\sqrt{R}}$ needs to be bounded for $\sqrt{T}$ consistency, while I require that $\frac{1}{\sqrt{R}}$ needs to be bounded for $\sqrt{T}$ consistency. The additional factor of $J$ in the fraction appears because all market shares are assumed to converge to 0.

2.3. Intuition with $J_i = 1$

To get a better sense of where the additional bias and variance terms come from, I now present an intuitive outline of the asymptotic normality proof with $J_i = 1$. The intuition for the general case is very similar. To simplify the notation, I drop $x_i$ as an argument from all functions. Since we are dealing with a GMM estimator, I show that the first order condition can be written as

$$\sqrt{T} \left( \hat{\theta} - \theta_0 \right) = \left( (\Gamma'W\Gamma)^{-1} \Gamma'W + o_p(1) \right) \sqrt{T} G_T(\theta_0, P_k, s^i) \left( \sqrt{T} \right) + o_p \left( \sqrt{T} \right),$$

where

$$G_T(\theta_0, P_k, s^i) = \frac{1}{T} \sum_{t=1}^{T} z_i^t \left( \xi_t(\theta_0, P_{Rt}, s_t) - \xi_t(\theta_0, P_{Rt}, s_t) \right).$$

Now consider $G_T(\theta_0, P_k, s^N)$ and write

$$G_T(\theta_0, P_k, s^N) = G_T(\theta_0, P_0, s) + G_T(\theta_0, P_R, s)$$

$$- G_T(\theta_0, P_0, s) + G_T(\theta_0, P_k, s^N) - G_T(\theta_0, P_R, s).$$

The first term would be the only term left if the integral was calculated exactly and if the true shares were observed. Since $E \left( z_i^t(\theta_0, P_{Rt}, s_t) \right) = 0$, the assumptions imply that $\sqrt{T} G_T(\theta_0, P_0, s)$ converges to a normally distributed random variable. Next write

$$G_T(\theta_0, P_k, s) = G_T(\theta_0, P_0, s)$$

$$= \frac{1}{T} \sum_{t=1}^{T} z_i^t \left( \xi_t(\theta_0, P_{Rt}, s_t) - \xi_t(\theta_0, P_{Rt}, s_t) \right).$$

This difference is not 0 because of the approximation error of the integral. To quantify the effect of the simulation error I use a Taylor expansion. I first rewrite $\xi_t(\theta_0, P_{Rt}, s_t)$ as a function of $\sigma_t(\theta, \xi_t, P_{Rt})$ and $\xi_t(\theta_0, P_{Rt}, s_t)$ as a function of $\sigma_t(\theta, \xi_t, P_{Rt})$, where $\sigma_t$ is defined in Eq. (3), using invertibility of this predicted share function. Since $\xi_t = \xi_t(\theta_0, P_{Rt}, s_t)$, we get $s_t = \sigma_t(\theta, \xi_t, P_{Rt}) = \sigma_t(\theta, \xi_t, P_{Rt}, s_t, P_{Rt})$ and thus

$$\xi_t(\theta_0, P_{Rt}, s_t) - \xi_t(\theta_0, P_{Rt}, s_t) = \sigma_t^{-1}(\theta, \sigma_t(\theta, \xi_t, P_{Rt}), P_{Rt}) - \sigma_t^{-1}(\theta, \sigma_t(\theta, \xi_t, P_{Rt}), P_{Rt}).$$

With this expression, I use a Taylor approximation to show that

$$\xi_t(\theta_0, P_{Rt}, s_t) = \xi_t(\theta_0, P_{Rt}, s_t) - H_{tR}^{-1} e_{Rt} - \frac{1}{2} H_{tR}^{-1} (e_{Rt})^2$$

$$+ \frac{1}{2} H_{tR}^{-1} \left( \frac{\partial e_{Rt}}{\partial \xi_t} \right) e_{Rt} + o_p \left( \sqrt{T} \right),$$

where $H_{tR}$ and $h_0$ constants,

$$e_{Rt} = \sigma_t(\theta, \xi_t, P_{Rt}) - \sigma_t(\theta, \xi_t, P_{Rt}),$$

and

$$\frac{\partial e_{Rt}}{\partial \xi_t} = \frac{\partial \sigma_t(\theta, \xi_t, P_{Rt})}{\partial \xi_t} - \frac{\partial \sigma_t(\theta, \xi_t, P_{Rt})}{\partial \xi_t}.$$

The first term on the right hand side of Eq. (7) result from the Taylor expansion while the fourth term arises because $P_{Rt}$ enters $\sigma_t^{-1}(\theta, \sigma_t(\theta, \xi_t, P_{Rt}), P_{Rt})$ both through $\sigma_t$ and as the last argument. Plugging this expansion back into (6) gives

$$\sqrt{T} \left( G_T(\theta_0, P_k, s) - G_T(\theta_0, P_0, s) \right) = - \frac{1}{\sqrt{T}} \sum_{t=1}^{T} z_i^t H_{tR}^{-1} e_{Rt}$$

$$+ \frac{1}{\sqrt{T} R} \sum_{t=1}^{T} z_i^t \left( H_{tR}^{-1} \left( \frac{\partial e_{Rt}}{\partial \xi_t} \right) e_{Rt} - \frac{1}{2} h_0 (e_{Rt})^2 \right) + o_p \left( \sqrt{T} \right).$$

The first term,

$$\frac{1}{\sqrt{T} R} \sum_{t=1}^{T} z_i^t H_{tR}^{-1} e_{Rt} = \frac{1}{\sqrt{T} R} \sum_{t=1}^{T} \sum_{r=1}^{R} z_i^t H_{tR}^{-1}$$

$$\times \left( v_{jt}(\theta, \xi_t, P_{Rt}) - \sigma_t(\theta, \xi_t, P_{Rt}) \right),$$

is $o_p \left( \frac{1}{\sqrt{T} R} \right)$ and converges to a normally distributed random variable when multiplied by $\sqrt{R}$. The second term does not have mean zero and it can be shown that the assumptions imply that

$$\frac{R}{\sqrt{T}} \frac{1}{\sqrt{T} R} \sum_{t=1}^{T} z_i^t \left( H_{tR}^{-1} \left( \frac{\partial e_{Rt}}{\partial \xi_t} \right) e_{Rt} - \frac{1}{2} h_0 (e_{Rt})^2 \right) \overset{p}{\to} \bar{\mu}.$$
Hence, it is $O_p\left(\frac{\sqrt{T}}{\sqrt{R}}\right)$ and converges in probability to a constant when multiplied by $\frac{R}{\sqrt{T}}$. Thus,

$$\sqrt{T} \left( G_t(\theta_0, P_R, s) - G_{\hat{\theta}}(\hat{\theta}_0, P_0, s) \right) = O_p\left(\frac{1}{\sqrt{R}}\right) + O_p\left(\frac{\sqrt{T}}{R}\right) + o_p\left(\frac{\sqrt{T}}{R}\right).$$

Similarly, using that $\xi_t(\theta_0, P_R, s_t^0) = \sigma_1^{-1}(\theta_0, s_t^0, P_R)$ and $\xi_t(\theta_0, P_R, s_t) = \sigma_t^{-1}(\theta_0, s_t, P_R)$ a Taylor approximation gives

$$\sqrt{T} \left( G_t(\theta_0, P_R, s) - G_{\hat{\theta}}(\hat{\theta}_0, P_0, s) \right) = \frac{1}{\sqrt{T}} \sum_{t=1}^T z_t^2 H_{0t}^{-1} (s_t^N - s_t) + \frac{1}{2} \frac{1}{\sqrt{T}} \sum_{t=1}^T z_t^2 H_{0t} (s_t^N - s_t)^2 + o_p\left(\frac{\sqrt{T}}{N}\right).$$

Again the assumptions imply that the first term on the right hand side converges to a normal distribution when multiplied by $\sqrt{N}$ and the second term converges to constant when multiplied by $\frac{\sqrt{T}}{\sqrt{N}}$.

Putting the previous results together yields the conclusions of Theorem 1. If $R$ and $N$ approach $\infty$ at a slower rate than $T$, the additional bias terms (second order terms in the Taylor expansion) dominate the additional variance terms (first order terms). Furthermore, the additional bias terms yield rate restrictions on $R$ and $N$ relative to $T$ whereas the first order terms converge to 0 as $R \to \infty$ and $N \to \infty$ at any rate. In the following Section 2 show how the leading bias term can be removed and how and the additional variance can be taken into account when calculating standard errors.

### 2.4 Bias and variance corrections

This section shows how the leading bias terms, namely the $O_p\left(\frac{\sqrt{T}}{\sqrt{R}}\right)$ term and the $O_p\left(\frac{\sqrt{T}}{R}\right)$ term in the expansion above, can be eliminated by using an analytic bias correction method. Similar methods have been suggested by Lee (1995), Arellano and Hahn (2007), and Kristensen and Salanié (2013) in related setups.

Furthermore, the leading additional variance terms, in particular the $O_p\left(\frac{\sqrt{T}}{\sqrt{R}}\right)$ term and the $O_p\left(\frac{\sqrt{T}}{R}\right)$ terms, can easily be taken into account when calculating standard errors.

#### 2.4.1 Analytic bias correction

Define the bias adjusted estimator as

$$\hat{\theta}_A \equiv \hat{\theta} - \left(\hat{\Gamma}^\prime W_{\hat{\Gamma}} \hat{\Gamma}\right)^{-1} \hat{\Gamma}^\prime W_{\hat{\Gamma}} \left(\frac{1}{\sqrt{T}} \hat{\mu}_1 + \frac{1}{N} \hat{\mu}_2\right)$$

where

$$\hat{\Gamma} = \frac{1}{T} \sum_{t=1}^T \frac{\partial}{\partial \theta} z_t^2 \xi_t \left(\hat{\theta}, P_{Rt}, s_t, x_t\right)$$

and expressions for $\hat{\mu}_1$ and $\hat{\mu}_2$, which are sample analogs of $\mu_1$ and $\mu_2$, are given in Section 3. Subtracting an estimator of $(\hat{\Gamma}^\prime W_{\hat{\Gamma}})^{-1} \hat{\Gamma}^\prime W_{\hat{\Gamma}} \left(\frac{1}{\sqrt{T}} \hat{\mu}_1 + \frac{1}{N} \hat{\mu}_2\right)$ from $\hat{\theta}$ eliminates the leading bias term from the asymptotic expansion, which is established by the following theorem.

### Theorem 2

Assume that Assumptions RC1–RC9 hold. Then

$$\sqrt{T} \left(\hat{\theta}_A - \theta_0\right) = \left(\left(\Gamma^\prime W \Gamma\right)^{-1} \Gamma^\prime W + o_p(1)\right) \times \left(Q_{1T} + \frac{1}{\sqrt{T}} Q_{2T} + \frac{1}{\sqrt{N}} Q_{3T} + o_p\left(\frac{\sqrt{T}}{R}\right) + o_p\left(\frac{\sqrt{T}}{N}\right)\right).$$

As opposed to the results of Theorem 1, now $\sqrt{T} \left(\hat{\theta}_A - \theta_0\right) \to N(0, V_1)$ as long as $\frac{\sqrt{T}}{R}$ is bounded.

#### 2.4.2 Variance correction

The variance of the estimator can be estimated by

$$\hat{V} = \left(\hat{\Gamma}^\prime W_{\hat{\Gamma}} \hat{\Gamma}^\prime W_{\hat{\Gamma}} \phi_1 + \frac{1}{\sqrt{T}} \phi_2 + \frac{1}{N} \phi_3\right) \hat{\Gamma} \left(\hat{\Gamma}^\prime W_{\hat{\Gamma}} \hat{\Gamma}^\prime W_{\hat{\Gamma}}\right)^{-1},$$

where $\phi_1$, $\phi_2$, and $\phi_3$ are estimators of $\phi_1$, $\phi_2$, and $\phi_3$ provided in Section 3. In this way the variance of the $O_p\left(\frac{\sqrt{T}}{\sqrt{R}}\right)$ term and the $O_p\left(\frac{\sqrt{T}}{R}\right)$ term is taken into account as well. In case a bias adjustment is used, $\hat{\theta}$ can be replaced by $\hat{\theta}_A$.

### 3. Bias and variance implementation in the BLP model

In this Section 1 explain that both the bias correction and the variance adjustment are easy to implement in the BLP model. In fact, most quantities needed to calculate these adjustments have already been calculated in other parts of the estimation procedure.

First recall that when the distribution of the random coefficients is $P_{Rt}$, the market shares predicted by the model are

$$\sigma_p(\theta, x_t, \xi_t, P_{Rt}) = \frac{1}{R} \sum_{t=1}^R v_p(\theta, x_t, \xi_t, v_t^1),$$

where

$$v_p(\theta, x_t, \xi_t, v_t) = \frac{\exp(x_t^\prime \beta + \xi_t^\prime + x_t^\prime \Sigma v)}{1 + \sum_{k=1}^K \exp(x_t^\prime \beta + \xi_t^\prime + x_t^\prime \Sigma v)}.$$}

Also $\sigma_1(\theta, x_t, \xi_t, P_{Rt})$ is the $1 \times 1$ vector with elements $\sigma_1(\theta, x_t, \xi_t, P_{Rt})$ and $v_t(\theta, x_t, \xi_t, v_t)$ is defined analogously. To calculate the variance of $\hat{\theta}_A$, even without any adjustments, one needs to calculate

$$\hat{\Gamma} = -\frac{1}{T} \sum_{t=1}^T z_t^2 \xi_t \left(\hat{\theta}, P_{Rt}, s_t^N, x_t\right).$$

This can be done by using the implicit function theorem which yields

$$\hat{\Gamma} = \frac{1}{T} \sum_{t=1}^T \frac{\partial \sigma_1(\hat{\theta}, x_t, \xi_t(\hat{\theta}, P_{Rt}, s_t^N, x_t), P_{Rt})}{\partial \xi_t} \left(\hat{\theta}, P_{Rt}, s_t^N, x_t\right)^{-1} \frac{\partial \sigma_1(\hat{\theta}, x_t, \xi_t(\hat{\theta}, P_{Rt}, s_t^N, x_t), P_{Rt})}{\partial \xi_t} \left(\hat{\theta}, P_{Rt}, s_t^N, x_t\right).$$

Let $\xi_t = \xi_t(\hat{\theta}, P_{Rt}, s_t^N, x_t)$. Notice that $\xi_t$ is needed in the estimation procedure when calculating the GMM objective function. When MPEC is used, $\xi_t$ are parameters. Hence, $\hat{\Gamma}$ mainly involves

---

8 An alternative to an analytic bias correction is a jackknife bias correction as in Kristensen and Salanié (2013). However, this estimator is computationally costly because one would have to solve the optimization problem several times. Furthermore, this procedure increases the variance of the estimator. Similarly, the panel jackknife bias correction suggested by Hahn and Newey (2004) is computationally very demanding in this setting.

9 The code is available on the author's homepage: http://www.ssc.wisc.edu/~jfreyberger.
derivatives of the predicted shares, which are easy to calculate. For example,
\[
\frac{\partial \sigma_i(\theta, x_t, \xi, P_R)}{\partial \xi_t} = \frac{1}{R} \sum_{r=1}^{R} \frac{\partial \nu_t(\theta, x_t, \xi, v_t)}{\partial \xi_t} = \frac{1}{R} \sum_{r=1}^{R} \text{diag} \left( \nu_t(\theta, x_t, \xi, v_t) \right)
\]
where \( \text{diag}(\nu_t) \) is a diagonal matrix with \( \nu_t \) on the diagonal. Similarly, one can calculate the derivative of the shares with respect to \( \theta \). Since these derivatives are already needed to calculate the unadjusted standard errors, they have been discussed in detail by Nevo (2000). Next define
\[
\hat{H}_t = \left( \frac{\partial \sigma_i(\hat{\theta}, x_t, \hat{\xi}_t, P_R)}{\partial \xi} \right)
\]
and
\[
\hat{I}_t = \sum_{k=1}^{k} \hat{H}_t^{-1} \left( \frac{\partial^2 \sigma_i(\hat{\theta}, x_t, \hat{\xi}_t, P_R)}{\partial \xi \partial \xi_k} \right) \hat{H}_t^{-1} \nu_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)
\]
where \( e_{jt} \) denotes the \( j \)th column of the \( J_t \times J_t \) identity matrix. To calculate \( \hat{I}_t \), we need the second derivative of the predicted shares.

It is easy to show that
\[
\frac{\partial^2 \sigma_i(\hat{\theta}, x_t, \hat{\xi}_t, v_t)}{\partial \xi \partial \xi_k} = \frac{1}{R} \sum_{r=1}^{R} \text{diag} \left( \frac{\partial \nu_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)}{\partial \xi_k} \right) - \left( \frac{\partial \nu_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)}{\partial \xi_k} \right)^{\prime} \nu_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)^{\prime}
\]
These are all the derivatives that we need to calculate. All adjustments are now simply functions of the previous quantities. So simplify the expression of the adjustments, define the demeaned version of \( \nu_t \) and its derivative by
\[
\hat{v}_t(\theta, x_t, \xi, v) = v_t(\theta, x_t, \xi, v) - \sigma_i(\theta, x_t, \xi, P_R)
\]
and
\[
\hat{\nu}_t(\theta, x_t, \xi, v) = \frac{\partial v_t(\theta, x_t, \xi, v)}{\partial \xi}
\]
as well as the demeaned individual purchasing decisions by
\[
\hat{s}_m = s_m - \frac{1}{N} \sum_{j=1}^{J} s_m.
\]
Now recall that the variance of the estimator can be estimated by
\[
\hat{\sigma}^2 = \left( \hat{\nu}^{\prime} \cdot \hat{W} \cdot \hat{\nu} \right)^{-1} \hat{\nu}^{\prime} \cdot \hat{W} \cdot \hat{\nu}
\]
Here
\[
\hat{\phi}_1 = \frac{1}{R} \sum_{r=1}^{R} \hat{s}_m \hat{s}_m^{\prime} \hat{z}_t, \quad \hat{\phi}_2 = \frac{1}{R} \sum_{r=1}^{R} \sum_{t=1}^{T} \hat{z}_t \hat{H}_t^{-1} \hat{v}_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t) \hat{v}_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)^{\prime} \hat{H}_t^{-1} \hat{v}_t, \quad \hat{\phi}_3 = \frac{1}{N} \sum_{j=1}^{J} \sum_{t=1}^{T} \hat{z}_t \hat{H}_t^{-1} \hat{s}_m \hat{s}_m^{\prime} \hat{H}_t^{-1} \hat{v}_t.
\]
We also have all ingredients to calculate the bias correction. To do so define
\[
\tilde{\hat{s}}_{jt} = \frac{1}{R} \sum_{r=1}^{R} \hat{v}_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t) \hat{v}_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)^{\prime}, \quad \tilde{\hat{s}}_{jt} = \frac{1}{N} \sum_{m=1}^{N} \hat{s}_m \hat{s}_m^{\prime}
\]
\[
\hat{\hat{c}}_t = \frac{1}{R} \sum_{r=1}^{R} \hat{H}_t^{-1} \left( \frac{\partial \nu_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)}{\partial \xi} \right) \hat{H}_t^{-1} \left( \frac{\partial \nu_t(\hat{\theta}, x_t, \hat{\xi}_t, v_t)}{\partial \xi} \right)^{\prime}
\]
All of the matrix are simply transformations of the first two derivatives of the predicted shares. The estimated bias terms are
\[
\hat{\hat{\mu}}_1 = \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{J} \hat{z}_t^{\prime} \left( \hat{e}_t \hat{c}_t - \frac{1}{2} \hat{s}_{jt} \right)
\]
and
\[
\hat{\hat{\mu}}_2 = - \frac{1}{2T} \sum_{t=1}^{T} \sum_{j=1}^{J} \hat{z}_t^{\prime} \hat{s}_{jt}.
\]

4. Monte Carlo simulations

In this section, I illustrate that the simulation error of the integral will affect the finite sample performance of the estimator because the usual GMM standard errors underestimate the true variance and the estimates are biased. I use the model described in Section 2. The setup for the Monte Carlo simulations is adapted from Dubé et al. (2012) with some changes to accommodate the asymptotics in the number of markets. A similar setup is also used by Skrainka and Judd (2011). The number of products is set to 4, I generate all shares using \( N = 3000 \) consumer, and I vary the number of markets, \( T \), and draws, \( R \), I include three additional product characteristics, which affect choices. One of the product characteristics does not vary across markets. These product characteristics, \( x_{kt} \) for \( k = 1, \ldots, 4 \), are distributed as \( TN(0, 1) \), where \( TN \) denotes the standard normal distribution truncated at \(-2 \) and \( 2 \) and \( x_{1jt} \) does not vary over markets. There is also a constant term, \( x_{0jt} = 1 \), for all \( j \). The unobserved product characteristics are \( \xi_j = \frac{1}{2} (\xi_j + \Delta \xi_j) \)
\[

\xi_j \sim TN(0, 1), \quad \Delta \xi_j \sim TN(0, 1), \quad j = 1, \ldots, 4, \quad t = 1, \ldots, T.
\]
Hence, also the unobserved product characteristics have a component which does not change over markets. I capture this part by using product dummies as parameters as in Nevo (2001).10 The price is
\[
\hat{p}_{jt} = \frac{1}{2} \left( 2 + 0.5 \xi_j + e_{jt} + 1.1 \sum_{k=1}^{4} x_{kjt} \right)
\]
where \( e_{jt} \sim TN(0, 1) \). There is a random coefficient on all product characteristics including price and the constant term. The random coefficient are distributed as follows
\[
\begin{pmatrix}
\beta_{10}^0 \\
\beta_{11}^0 \\
\beta_{12}^0 \\
\beta_{13}^0 \\
\beta_{14}^0 \\
\beta_{15}^0 \\
\beta_{16}^0 \\
\beta_{17}^0 \\
\beta_{18}^0 \\
\beta_{19}^0 \\
\beta_{110}^0
\end{pmatrix} \sim N \left( \begin{pmatrix}
3 \\
1.5 \\
1.5 \\
1.5 \\
1.5 \\
1.5 \\
1.5 \\
1.5 \\
1.5 \\
1.5 \\
1.5
\end{pmatrix}, \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix} \right)
\]

10 See footnote 5 for a discussion on using product dummies as parameters.
Here $\beta^s_0$ and $\beta^p_0$ denote the random coefficients on $x_{ikt}$ and $p_{it}$, respectively. Allowing for either correlated product characteristics or correlated random coefficients does not change the simulation results much. For each product $j$ in market $t$, I generate 3 instruments

$Z_{ikt} = a_{ikt} + 0.25 \left( e_g + 1.1 \sum_{k=1}^{4} x_{ikt} \right)$.

$I = 1, \ldots, J, \ j = 1, \ldots, 4, \ t = 1, \ldots, T$,

where $a_{ikt} \sim U(0, 1)$. Next to $Z_{ikt}$, $x_{ikt}$, $J_{ikt}$, and product dummies, I also use $x_{ikt}$ and $Z_{ikt}$ for all $I$, $x_{ikt}$, $J_{ikt}$, and $x_{ikt}$ as well as $x_{ikt}$, $J_{ikt}$, $x_{ikt}$ and $x_{ikt}$ as instruments. Just as Su and Judd (2012) and Dubé et al. (2012) I use a reduced form process for the price and the instruments instead of generating them from a structural model. Consequently, the instruments here might be stronger than those commonly found in applications.

I make use of the Mathematical Programming with Equilibrium Constraints (MPEC) approach (see Su and Judd, 2012 and Dubé et al., 2012) to estimate the BLP model, where I use KNITRO (Version 8.1) to solve the constrained optimization problem. I supply the solver with analytic gradients for the objective function and the constraints but not with the Jacobian. I use the same tolerance levels for the constraints and the objective function as Dubé et al. (2012). I also use a similar method as Dubé et al. (2012) to obtain 5 different starting values for each simulated data set. That is, I first draw values for the standard deviations of the random coefficients from a $U(0, 1)$ distribution. I then obtain the starting values for the means (and the other parameters for the MPEC procedure) with a two stage least squares procedure taking the variances as given. The starting values correspond to the minimum of the objective function (10) for a given value of the variances. As a consequence, all starting values are feasible. In the majority of the simulated data sets, even when $R = 50$, most starting values yield the same minimum of the objective function. For example, when $R = 50$ and $T = 400$ all five starting values yield the same minimum in over 60% of the simulated data sets. Moreover, when $R$ increases it is more likely that all starting values converge to the same solution.

Below I investigate the actual coverage rate of a 95% nominal confidence interval for $\beta^p_0 = E(\beta^p_0) = -2$ using bias correction methods as well as standard errors with and without correcting for the approximation errors. The usual GMM standard errors are estimated using

$\hat{V}_i = \left( \hat{W}^{\prime} \hat{W} \right)^{-1} \hat{W}^{\prime} \phi_i \hat{W} \left( \hat{W}^{\prime} \hat{W} \right)^{-1}$

and the adjusted standard errors are calculated using

$\hat{V}_i = \left( \hat{W}^{\prime} \hat{W} \right)^{-1} \hat{W}^{\prime} \left( \phi_1 + \frac{1}{R} \phi_2 + \frac{1}{R^2} \phi_3 \right) \hat{W} \left( \hat{W}^{\prime} \hat{W} \right)^{-1}$.

I also compare the mean lengths of the confidence intervals obtained from simulations with and without variance corrections. I simulate the data with 50,000 draws from the joint distribution of random coefficients. I treat these 50,000 draws as the true distribution and then take 50–800 random draws from this distribution depending on the setup. I make use of 50–800 markets. I use different draws to approximate the integral in different markets. Hence, with 100 markets and 800 draws I sample in total 80,000 times from the true distribution. The computational costs mainly depend on the number of draws that are used to evaluate each integral, which is 800 in this case. All coverage rates are based on 1000 Monte Carlo iterations, i.e. 1000 simulated data sets.

Table 1 shows that the number of draws affects the actual coverage rate of a nominal 95% confidence interval if the usual GMM asymptotic distribution is employed. For example with 800 markets the actual coverage rate is only 74.5% with 50 draws, while it increases to 91.1% with 800 draws. Using the bias corrected estimator leads to an improvement in these coverage rates. However, in general there is still a large difference between using a small number and large number of draws. For instance, with 800 markets and 50 draws one obtains a coverage rate of 85.3%, while 800 draws yield a coverage rate of 94.8%. The same holds when simulation adjusted standard errors but no bias adjustment is used. In case one uses both the analytical bias adjustment and the standard error adjustment, the coverage rate is very close to 95% even with a small number of draws. For example with 800 markets and 50 draws the coverage rate is 92.0%. Therefore, as mentioned by Skrainka and Judd (2011), the usual GMM standard errors are too tight when the number of draws is small, but the adjusted standard seems to be of the right magnitude. The costs of the improved coverage rates are wider confidence intervals. Table 2 shows the mean length of confidence intervals with adjusted standard errors divided by the mean length with unadjusted standard errors. The means are taken over the 1000 simulated data sets. When the number of draws is 50 the adjusted confidence intervals are up to 25% larger. When the number of draws increases, this ratio decreases. With 800 draws, the difference is less than 10%. Thus, corrected standard errors only have a large effect on coverage rates and the lengths of the confidence intervals when the number of simulations $R$ is small.

Table 3 shows the finite sample bias with and without bias correction. The finite sample bias of the GMM point estimates decreases as $R$ increases and as $T$ increases. Especially if $T$ is small, the finite sample bias is still substantial even when $R$ is large, because this nonlinear estimator is biased in finite samples even if $R = \infty$. The bias correction reduces the finite sample bias significantly. For the bias corrected point estimates, there is no significant difference in the bias when $R$ is large and when $R$ is small (especially when $T$ is large), which suggest that the bias correction eliminates most of the bias due to the simulation error. The remaining bias is due to the nonlinear objective function.

Finally, Table 4 shows the ratio of the mean squared error of $\hat{\beta}_0$ when the same and different draws are used to approximate the integral in different markets. When the number of draws is small, the mean squared error with the same draws is up to 37.2% larger (with 400 markets). As the number of draws increases, this
Table 1
Coverage rates of 95% confidence intervals for $\beta^p$.

<table>
<thead>
<tr>
<th></th>
<th>50 draws</th>
<th>100 draws</th>
<th>200 draws</th>
<th>400 draws</th>
<th>800 draws</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 markets</td>
<td>0.784</td>
<td>0.860</td>
<td>0.849</td>
<td>0.856</td>
<td>0.904</td>
</tr>
<tr>
<td>100 markets</td>
<td>0.786</td>
<td>0.839</td>
<td>0.889</td>
<td>0.890</td>
<td>0.885</td>
</tr>
<tr>
<td>200 markets</td>
<td>0.798</td>
<td>0.849</td>
<td>0.885</td>
<td>0.885</td>
<td>0.915</td>
</tr>
<tr>
<td>400 markets</td>
<td>0.797</td>
<td>0.851</td>
<td>0.896</td>
<td>0.898</td>
<td>0.909</td>
</tr>
<tr>
<td>800 markets</td>
<td>0.745</td>
<td>0.847</td>
<td>0.883</td>
<td>0.903</td>
<td>0.911</td>
</tr>
<tr>
<td>Bias correct.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 markets</td>
<td>0.833</td>
<td>0.908</td>
<td>0.893</td>
<td>0.893</td>
<td>0.929</td>
</tr>
<tr>
<td>100 markets</td>
<td>0.869</td>
<td>0.892</td>
<td>0.937</td>
<td>0.934</td>
<td>0.926</td>
</tr>
<tr>
<td>200 markets</td>
<td>0.889</td>
<td>0.909</td>
<td>0.940</td>
<td>0.927</td>
<td>0.949</td>
</tr>
<tr>
<td>400 markets</td>
<td>0.892</td>
<td>0.930</td>
<td>0.943</td>
<td>0.946</td>
<td>0.947</td>
</tr>
<tr>
<td>800 markets</td>
<td>0.853</td>
<td>0.909</td>
<td>0.932</td>
<td>0.951</td>
<td>0.948</td>
</tr>
<tr>
<td>GMM estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 markets</td>
<td>0.870</td>
<td>0.908</td>
<td>0.893</td>
<td>0.891</td>
<td>0.924</td>
</tr>
<tr>
<td>100 markets</td>
<td>0.865</td>
<td>0.891</td>
<td>0.925</td>
<td>0.928</td>
<td>0.919</td>
</tr>
<tr>
<td>200 markets</td>
<td>0.888</td>
<td>0.893</td>
<td>0.923</td>
<td>0.911</td>
<td>0.933</td>
</tr>
<tr>
<td>400 markets</td>
<td>0.865</td>
<td>0.902</td>
<td>0.921</td>
<td>0.924</td>
<td>0.933</td>
</tr>
<tr>
<td>800 markets</td>
<td>0.821</td>
<td>0.893</td>
<td>0.907</td>
<td>0.932</td>
<td>0.930</td>
</tr>
<tr>
<td>Bias correct.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 markets</td>
<td>0.911</td>
<td>0.943</td>
<td>0.920</td>
<td>0.913</td>
<td>0.942</td>
</tr>
<tr>
<td>100 markets</td>
<td>0.924</td>
<td>0.939</td>
<td>0.957</td>
<td>0.959</td>
<td>0.937</td>
</tr>
<tr>
<td>200 markets</td>
<td>0.939</td>
<td>0.948</td>
<td>0.958</td>
<td>0.947</td>
<td>0.965</td>
</tr>
<tr>
<td>400 markets</td>
<td>0.943</td>
<td>0.945</td>
<td>0.960</td>
<td>0.963</td>
<td>0.963</td>
</tr>
<tr>
<td>800 markets</td>
<td>0.920</td>
<td>0.947</td>
<td>0.954</td>
<td>0.966</td>
<td>0.964</td>
</tr>
</tbody>
</table>

The nominal coverage rate is 0.95 and the number of Monte Carlo iterations is 1000. The true value of $\beta^p$ is $-2$. If the actual coverage rate is 95%, the standard error with 1,000 simulations is around 0.0069. If the actual coverage rate is 80%, the standard error increases to 0.0126.

Table 2
Ratio of mean length of unadjusted and adjusted confidence intervals for $\beta^p$.

<table>
<thead>
<tr>
<th></th>
<th>50 draws</th>
<th>100 draws</th>
<th>200 draws</th>
<th>400 draws</th>
<th>800 draws</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 markets</td>
<td>1.251</td>
<td>1.174</td>
<td>1.129</td>
<td>1.110</td>
<td>1.089</td>
</tr>
<tr>
<td>100 markets</td>
<td>1.239</td>
<td>1.179</td>
<td>1.134</td>
<td>1.112</td>
<td>1.093</td>
</tr>
<tr>
<td>200 markets</td>
<td>1.246</td>
<td>1.178</td>
<td>1.134</td>
<td>1.108</td>
<td>1.095</td>
</tr>
<tr>
<td>400 markets</td>
<td>1.244</td>
<td>1.183</td>
<td>1.140</td>
<td>1.113</td>
<td>1.097</td>
</tr>
<tr>
<td>800 markets</td>
<td>1.240</td>
<td>1.180</td>
<td>1.137</td>
<td>1.112</td>
<td>1.098</td>
</tr>
</tbody>
</table>

Table 3
Finite sample bias of $\hat{\beta}^p$.

<table>
<thead>
<tr>
<th></th>
<th>50 draws</th>
<th>100 draws</th>
<th>200 draws</th>
<th>400 draws</th>
<th>800 draws</th>
</tr>
</thead>
<tbody>
<tr>
<td>GMM estimates</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 markets</td>
<td>-0.131</td>
<td>-0.110</td>
<td>-0.127</td>
<td>-0.129</td>
<td>-0.126</td>
</tr>
<tr>
<td>100 markets</td>
<td>-0.112</td>
<td>-0.101</td>
<td>-0.095</td>
<td>-0.097</td>
<td>-0.100</td>
</tr>
<tr>
<td>200 markets</td>
<td>-0.091</td>
<td>-0.083</td>
<td>-0.080</td>
<td>-0.081</td>
<td>-0.085</td>
</tr>
<tr>
<td>400 markets</td>
<td>-0.087</td>
<td>-0.070</td>
<td>-0.061</td>
<td>-0.065</td>
<td>-0.062</td>
</tr>
<tr>
<td>800 markets</td>
<td>-0.078</td>
<td>-0.060</td>
<td>-0.056</td>
<td>-0.054</td>
<td>-0.052</td>
</tr>
<tr>
<td>Bias correct.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50 markets</td>
<td>-0.076</td>
<td>-0.065</td>
<td>-0.087</td>
<td>-0.092</td>
<td>-0.091</td>
</tr>
<tr>
<td>100 markets</td>
<td>-0.055</td>
<td>-0.052</td>
<td>-0.052</td>
<td>-0.057</td>
<td>-0.063</td>
</tr>
<tr>
<td>200 markets</td>
<td>-0.022</td>
<td>-0.032</td>
<td>-0.036</td>
<td>-0.040</td>
<td>-0.045</td>
</tr>
<tr>
<td>400 markets</td>
<td>-0.019</td>
<td>-0.014</td>
<td>-0.014</td>
<td>-0.021</td>
<td>-0.021</td>
</tr>
<tr>
<td>800 markets</td>
<td>-0.013</td>
<td>-0.006</td>
<td>-0.009</td>
<td>-0.010</td>
<td>-0.010</td>
</tr>
</tbody>
</table>

Table 4
Ratio of mean squared error of $\hat{\beta}^p$ with same and different draws.

<table>
<thead>
<tr>
<th></th>
<th>50 draws</th>
<th>100 draws</th>
<th>200 draws</th>
<th>400 draws</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 markets</td>
<td>0.937</td>
<td>1.134</td>
<td>1.022</td>
<td>1.031</td>
</tr>
<tr>
<td>100 markets</td>
<td>1.099</td>
<td>0.995</td>
<td>1.116</td>
<td>1.121</td>
</tr>
<tr>
<td>200 markets</td>
<td>1.363</td>
<td>1.162</td>
<td>1.160</td>
<td>0.977</td>
</tr>
<tr>
<td>400 markets</td>
<td>1.372</td>
<td>1.203</td>
<td>1.204</td>
<td>1.056</td>
</tr>
</tbody>
</table>

The ratio converges to 1. Table A1 in the supplementary appendix (see Appendix A) presents additional simulation results when the same draws are used, which show that the adjustments are also important in this case. The unadjusted confidence intervals generally have a lower coverage rate than the ones obtained with different draws, especially when $R$ is small relative to $T$. The adjustments again yield coverage rates close to 95% in all cases. When the same draws are used the adjustments are more costly. In particular, the ratio of the mean length of the adjusted and unadjusted confidence intervals now depends on $T$ and $R$ (as opposed to only $R$ with different draws) and adjusted confidence intervals are up to 72.4% larger. As a consequence, adjusted confidence intervals are larger when the same draws are used compared to using different draws. For example, when $T = 400$ and $R = 50$, the mean length of adjusted confidence intervals with
the same draws is 24.5% larger. These results are all in line with the theoretical findings in this paper and verify that different draws should be used to calculate the integrals in different markets.

Clearly, many parameter choices drive the results in this Monte Carlo study. For example, a low variance of the error term (relative to the variance of the product characteristics) or strong instruments yield more precise estimates for a given number of markets. Furthermore, the effect of the number of draws depends on the variance of the random coefficients relative to the variance of the product characteristics. A high variance of the random coefficient implies that we need many draws to eliminate the effect of the second term of the asymptotic expansion. It is therefore hard to give a general guideline of how many draws (or how many markets) suffice to obtain satisfactory results. The simulation results, however, demonstrate that practitioners should use the bias correction and the adjusted standard errors in application. If the number of draws is sufficiently large, the bias correction is small and the adjusted standard errors will be very close to the GMM standard errors. If the number of draws is small, the simulation error will affect the finite sample performance of the estimator and using the usual GMM asymptotic distribution yields biased estimators and underestimation of the true variance. Since a large number of draws improves the precision of the initial estimator, which is in turn used to calculate the bias correction, the number of draws should be as large as possible, subject to computational constraints and data availability.

5. Conclusion

This paper develops asymptotic theory for estimated parameters in differentiated product demand systems with a small number of products and a large number of markets T. The asymptotic theory takes into account that the predicted market shares are approximated by Monte Carlo integration with R draws and that the observed market shares are approximated from a sample of N consumers. Both approximations affect the asymptotic distribution, because they both lead to a bias and a variance term in the asymptotic expansion of the estimator. I show that when R and N do not increase faster than the number of markets, the bias terms dominate the variance terms. In this case, the asymptotic distribution might not be centered at 0 and standard confidence intervals do not have the right size, even asymptotically. These findings differ from the setup with a large number of products where the variance term always dominates.

I propose both bias corrections and variance adjustments in order to take the approximation errors into account. I then demonstrate with Monte Carlo simulations that these adjustments, which are easy to compute, should be used in applications. In particular, I show that if the number of draws and the number of consumers are sufficiently large, the bias correction is small and the adjusted standard errors will be very close to the GMM standard errors. However, if the number of draws or the number of consumers is small, the approximations will affect the finite sample performance of the estimator and using the usual GMM asymptotic distribution yields biased estimators and underestimation of the true variance. As a consequence the coverage rate of confidence intervals can be significantly below the nominal rate. The estimates and confidence intervals with the adjustments do not suffer from these issues.

Appendix A. Useful lemmas

I use the following lemmas to prove the main results. The proofs are in the supplementary appendix (see Appendix A).

Lemma A.1. Let \( f(x, \theta, v) : \mathcal{X} \times \Theta \times \mathbb{R}^d \to [-M_1, M_2] \subset \mathbb{R} \) be a continuously differentiable function in all arguments where \( \mathcal{X} \) is a compact subset of \( \mathbb{R}^{d_x} \) and \( \Theta \) is a compact subset of \( \mathbb{R}^{d_\theta} \). Let \( v_1, \ldots, v_R \) be iid draws from \( P_1 \in \mathbb{P} \). Let \( x_i \in \mathcal{X} \) denote the (random) data. Let \( \tilde{P} = \{P^1, \ldots, P^R\} \) be a finite set of distributions. Assume that

(i) \( \ln(T)/R(T) \to 0 \) as \( T \to \infty \),

(ii) \( v_{it} \) and \( x_i \) are independent, and

(iii) Any random variable \( v_i \) with distribution \( P_i \) satisfies the following: \( v_i = g(a_i, w_i) \) where \( g : \mathcal{A} \times \mathbb{R}^{d_w} \to \mathbb{R}^{d_v} \) is continuous in both arguments, \( a_i \) is a vector of constants and \( \mathcal{A} \subset \mathbb{R}^{d_a} \) is a compact set, \( w_i \) is a random vector and the distribution of each element of \( w_i \) either (a) is in \( \tilde{P} \) and is the same for all \( t \) or (b) the support is a subset of \( \{M_1, M_2\} \) for some constants \( M_1 \) and \( M_2 \). Then

\[
\sup_{\theta \in \Theta} \max_{1 \leq t \leq T} \left| \frac{1}{R} \sum_{r=1}^{R} f(x_i, \theta, v_{ir}) - \int f(x_i, \theta, v) dP_i(v) \right| \overset{p}{\to} 0
\]

as \( T \to \infty \).

Lemma A.2. Let \( Y_1, Y_2, \ldots \) be independent random variables with mean 0. Let \( K \) be an even integer. Then

\[
E \left( \left| \frac{1}{R} \sum_{r=1}^{R} Y_r \right|^K \right) \leq A_K R^{(K/2-1)} \sum_{r=1}^{R} E \left( |Y_r|^K \right)
\]

where \( A_K \) is a universal constant depending only on \( K \).

Lemma A.3. Suppose that \( (v^{(i)}_1, \ldots, v^{(i)}_R, x_i) \) with \( t = 1, \ldots, T \) are random vectors such that \( v^{(i)}_r \) are iid across \( r \) and independent of \( x_i \). Let

\[
y_{ir} = s(x_i) \left( \frac{1}{R} \sum_{r=1}^{R} p \left( v^{(i)}_r, x_i \right) \right)^{m_1} \left( \frac{1}{R} \sum_{r=1}^{R} q \left( v^{(i)}_r, x_i \right) \right)^{m_2}
\]

where \( m_1 \) and \( m_2 \) are nonnegative integers and \( s, p, q \) are measurable functions such that

\[
E \left( p \left( v^{(i)}_r, x_i \right) | x_i \right) = E \left( q \left( v^{(i)}_r, x_i \right) | x_i \right) = 0
\]

for all \( t \). Also assume that if \( m_1 > 0 \) and \( m_2 > 0 \), then for some \( a \) and \( b \) satisfying \( \frac{1}{a} + \frac{1}{b} = 1 \) it holds that for some finite \( M \)

\[
E \left( \left| s(x_i) \right|^{2a} p \left( v^{(i)}_r, x_i \right)^{2m_1} \right) \leq M
\]

and

\[
E \left( \left| s(x_i) \right|^{2a} q \left( v^{(i)}_r, x_i \right)^{2m_2} \right) \leq M.
\]

If \( m_1 > 0 \) and \( m_2 = 0 \) assume instead that

\[
E \left( \left| s(x_i) \right|^{2a} p \left( v^{(i)}_r, x_i \right)^{2m_1} \right) \leq M.
\]

Then

\[
\frac{1}{T} \sum_{t=1}^{T} |y_{ir}| = O_p \left( R^{-\left(m_1+m_2\right)/2} \right).
\]

Lemma A.4. Suppose \( F : \mathbb{R}^d \to \mathbb{R} \) is a \( k + 1 \) times continuously differentiable function on an open convex set \( S \subset \mathbb{R}^d \). For any multi-index \( \alpha \in \mathbb{N}^d_+ \), let \( \left| \alpha \right| = \alpha_1 + \cdots + \alpha_d \) and \( \alpha! = \alpha_1! \cdots \alpha_d! \). Furthermore, for \( x \in \mathbb{R}^d \) let

\[
x^{(\alpha)} = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}.
\]
and
\[ \frac{\partial^{[\omega]} F(x)}{\partial x^{[\omega]}} = \frac{\partial^{[\omega]} F(x)}{\partial x^{[\omega]}} h^{[\omega]} \cdot \partial x^{[\omega]} \cdot \partial x^{[\omega]} \ldots \partial x^{[\omega]} \partial x^{[\omega]} . \]

If \( a \in S \) and \( a + h \in S \), then for some \( c \in (0, 1) \)
\[ F(a + h) = \sum_{a \in \mathbb{N}^{[\omega]}} \frac{1}{a!} \left( \frac{\partial^{[\omega]} F(x)}{\partial x^{[\omega]}} |_{x = a} \right) h^{[\omega]} \]
\[ + \sum_{a \in \mathbb{N}^{[\omega]}} \frac{1}{a!} \left( \frac{\partial^{[\omega]} F(x)}{\partial x^{[\omega]}} |_{x = a + ch} \right) h^{[\omega]} . \]

Appendix B. A general model

In this section I present a general model, high level assumptions, and consistency and asymptotic normality results. The BLP is a special case of this general setup and I prove Theorem 1 by verifying the assumptions in this section. I make use of the following notation. I denote the norm of a \( a \times b \) matrix \( \| A \| = tr(AA^T)^{1/2} \) and I denote a neighborhood of a vector \( a_0 \in \mathbb{R}^d \) by \( N_\epsilon(a_0) = \{ a \in \mathbb{R}^d : \| a - a_0 \| \leq \epsilon \} \).

Let \( x_t \in \mathbb{R}^{d_1} \), \( \xi_t \in \mathbb{R}^{d_2} \), and \( \Theta \subset \mathbb{R}^{d_3} \) be defined as in Section 2. Suppose there exists a \( d_t \) dimensional random vector \( v \) with distribution function \( P_v \) which may change over markets. I assumed that \( P_v \) is known, but \( P_v \in \mathcal{P} \) where \( \mathcal{P} \) is a space of probability distributions, which is restricted in the assumptions that follow. For given product characteristics \((x_t, \xi_t)\), a parameter value \( \theta \), and a distribution \( P_t \), the \( J_t \times 1 \) vector of market shares in market \( t \) predicted by the model is denoted by \( \sigma_t(\theta, x_t, \xi_t, P_t) \).

Define the function \( v_t : \mathbb{R}^{d_1} \times \Theta \times \mathbb{R}^{d_2} \rightarrow \mathbb{R}^{l} \). Assume that the predicted shares have the form
\[ \sigma_t(\theta, x_t, \xi_t, P_t) = \int v_t(\theta, x_t, \xi_t, v) dP_v(v), \quad \forall t = 1, \ldots, T . \]

The \( J_t \times 1 \) vector of true market shares in market \( t \) is denoted by \( s_t \) and I assume that
\[ s_t = \int v_t(\theta_0, x_t, \xi_t, v) dP_{\theta_0}(v), \quad \forall t = 1, \ldots, T \quad (8) \]
for some \( \theta_0 \) and \( P_{\theta_0} \in \mathcal{P} \) which is analogous to the discussion in Section 2 but without specific functional form or distributional assumptions. In particular, in Section 2
\[ v_t(\theta, x_t, \xi_t, v) = \frac{\exp(x_t^\beta + \xi_t^\gamma + x_t^\beta \Sigma v)}{1 + \sum_{k=1}^{K} \exp(x_t^\beta + \xi_t^\gamma + x_t^\beta \Sigma v)} . \]
A different functional form for \( v_t \) arises for example if \( \epsilon_{ijt} \) in (1) is not extreme value distributed.

To simplify notation, I suppress \( x_t \) and refer to \( v_t(\theta, x_t, \xi_t, v) \) as \( v_t(\theta, \xi_t, v) \) and to \( \sigma_t(\theta, x_t, \xi_t, P_t) \) as \( \sigma_t(\theta, \xi_t, P_t) \). The \( j \)-th element of \( \sigma_t(\theta, \xi_t, P_t) \) is denoted by \( \sigma_{jt}(\theta, \xi_t, P_t) \). To state the moment conditions and the estimator of \( \theta_0 \), I make the following assumptions.

Assumption A.1. For any pair \((P_t, \theta) \in \mathcal{P} \times \Theta \) and vector \((m_t, s_t) \in \mathcal{M} \times (0, 1)^d \) there exists a unique solution \( \xi_t \) to \( s_t - \sigma_t(\theta, \xi_t, P_t) = 0 \). This solution is denoted by \( \xi_t(\theta, P_t, s_t, x_t) \) and abbreviated by \( \xi_t(\theta, P_t, s_t) \).

Next define the function
\[ G_t(\theta, P_t) \equiv E(z_t \xi_t(\theta, P_t, s_t)), \]
and the sample moment
\[ G_t(\theta, P_t, s_t) \equiv \frac{1}{T} \sum_{t=1}^{T} z_t \xi_t(\theta, P_t, s_t) \]
where \( z_t \in \mathbb{R}^{d_1} \) is a vector of instrumental variables. The moment conditions of the model are
\[ E(z_t \xi_t(\theta_0, P_{\theta_0}, s_t)) = 0. \quad (9) \]

The function \( \sigma_t(\theta, \xi_t, P_{\theta_0}) \) involves an integral, which is approximated using the empirical probability measure \( P_{\theta_0} \) sample \( v_t, \ldots, v_{t_T} \) from \( P_{\theta_0} \). We may also not observe the true market shares \( s_t \) but instead an approximation from a sample of \( N \) consumers. That is, for \( n = 1, \ldots, N \) and \( t = 1, \ldots, T \), we observe \( s_m \in [0, 1]^l \) with \( E(s_m) = s_t \) and \( \sum_{m=1}^{N} s_m \leq 1 \). Define the observed approximated market shares by \( s_t^{(N)} = \frac{1}{N} \sum_{m=1}^{N} s_m \).

I assume that the number of draws, \( R \), and the number of consumers, \( N \), are functions of \( T \) and all limits are taken as \( T \rightarrow \infty \), now define the estimator
\[ \hat{\theta} \equiv \arg \min_{\theta \in \mathcal{P}} \left( \frac{1}{T} \sum_{t=1}^{T} z_t \xi_t(\theta, P_{\theta_0}, s_t^{N}) \right)^T \times W_t \left( \frac{1}{T} \sum_{t=1}^{T} z_t \xi_t(\theta, P_{\theta_0}, s_t^{N}) \right) . \quad (10) \]

The estimator \( \hat{\theta} \) is the one used in practice and its asymptotic properties are analyzed here.

B.1. Consistency

I first provide general assumptions, which are sufficient conditions for consistency.

Assumption A.2. The market shares are generated by (8) and moment conditions (9) hold.

Assumption A.3. \( J_t \leq J \) for some integer \( J \).

Assumption A.4. The data \((z_t, x_t, p_t, s_t)_{t=1}^{T}\) are independent across markets.

Assumption A.5. The matrix \( \frac{1}{N} \sum_{t=1}^{T} z_t \xi_t \) has full rank and is stochastically bounded, i.e. for all \( \epsilon > 0 \) there exists an \( M(\epsilon) \) such that \( \Pr\left( \| \frac{1}{N} \sum_{t=1}^{T} z_t \xi_t \| > M(\epsilon) \right < \epsilon . \)

Assumption A.6. \( E(\xi_t(\theta_0, P_{\theta_0}, s_t) | z_t) \leq M \) for all \( t = 1, \ldots, T \) and \( j = 1, \ldots, J_t \).

Assumption A.7. As \( T \rightarrow \infty \)
\[ \sup_{\theta_0} \max_{1 \leq t \leq T, 1 \leq j \leq J_t} | \sigma_{jt}(\theta_0, \xi_t(\theta_0, P_{\theta_0}, s_t), P_{\theta_0}) | \rightarrow 0 \]
and
\[ \max_{1 \leq t \leq T, 1 \leq j \leq J_t} | s_t^{(N)} - s_t | \rightarrow 0 . \]

Assumption A.8. For all \( \delta > 0 \), there exists \( C(\delta) > 0 \) such that for all \( t = 1, \ldots, T \),
\[ \lim_{t \rightarrow \infty} \Pr\left( \inf_{\theta \in \mathcal{P}} \left\{ \min_{1 \leq t \leq T, 1 \leq j \leq J_t} | \sigma_t(\theta, \xi_t, P_{\theta_0}) - \sigma_t(\theta_0, \xi_t, P_{\theta_0}) | \right\} \geq C(\delta) \right) = 1 . \]
Assumption A.9. For all $\delta > 0$, there exists $C(\delta) > 0$ such that
\[
\lim_{T \to \infty} \Pr \left( \inf_{\theta \in \Theta_0} \| G_T(\theta, P_k, s^N) - G_T(\theta, P_0, s) \| \geq C(\delta) \right) = 1.
\]

I assume that the number of products in each market is bounded in Assumption A.3 and that we observe independent markets in Assumption A.4. Assumptions A.5 and A.6 state that the matrix of instruments has full rank, is stochastically bounded, and that the unobserved product characteristics have a bounded second conditional moment.

Assumptions A.7 and A.8 are needed in the proof because it is required that
\[
\sup_{\delta \in \Theta} \| G_T(\theta, P_k, s^N) - G_T(\theta, P_0, s) \| = 0(1).
\]

A sufficient condition for this to hold is that $\xi(\theta, P_k, s^N) - \xi(\theta, P_0, s)$ converges to 0 in probability uniformly over $\theta$ and $t$. Since there is no closed form expression for $\xi$, I assume instead that the market shares generated by the model, with the true and the approximated distribution, as well as the true and approximated market shares are uniformly close (Assumption A.7) and that this would be violated if $\xi(\theta, P_0, s)$ was not close to $\xi(\theta, P_K, s)$ (Assumption A.8).

Assumption A.8 says that asymptotically the $\xi$ that sets the model's predictions for shares equal to the actual shares can be distinguished from other values of $\xi$. Assumption A.7 implies that $R(T) \to \infty$ and $N(T) \to \infty$ as $T \to \infty$. All assumptions can be verified under low level condition in the BLP model as explained in Section 2.1. Finally, Assumption A.9 states that $\theta_0$ is identified from the moment conditions.

These assumptions imply the following theorem.

Theorem A.1. Suppose that Assumptions A.1–A.9 hold. Then $\hat{\theta} \stackrel{p}{\to} \theta$ as $T \to \infty$.

Proof. The proof consists of two parts. In the first part I show that an estimator defined as any sequence that satisfies $\| G_T(\hat{\theta}, P_0, s) \| = \inf_{\delta \in \Theta} \| G_T(\delta, P_0, s) \| + o_p(1)$ is a consistent estimator for $\theta$. To do so notice that for any $\delta > 0$,
\[
\Pr \left( \| \hat{\theta} - \theta_0 \| \geq \delta \right)
\leq \Pr \left( \| \hat{\theta} - \theta_0 \| \geq \delta, \| G_T(\theta_0, P_0, s) - G_T(\hat{\theta}, P_0, s) \| \geq C(\delta) \right)
+ \Pr \left( \| G_T(\theta_0, P_0, s) - G_T(\hat{\theta}, P_0, s) \| \geq C(\delta) \right)
\leq \Pr \left( \inf_{\delta \in \Theta_0} \| G_T(\theta_0, P_0, s) - G_T(\theta, P_0, s) \| \geq C(\delta) \right).
\]

The second term on the right hand side converges to 0 by Assumption A.9. For the first term notice that Assumptions A.4–A.6, and Kolmogorov's first law of large numbers implies that $\| G_T(\theta_0, P_0, s) \| = o_p(1)$. Thus
\[
\| G_T(\theta_0, P_0, s) - G_T(\hat{\theta}, P_0, s) \| \leq 2 \| G_T(\theta_0, P_0, s) \| + o_p(1) = o_p(1).
\]

Therefore, $\hat{\theta} \stackrel{p}{\to} \theta_0$.

In the second part I show that $\sup_{\delta \in \Theta} \| G_T(\theta, P_k, s^N) - G_T(\theta, P_0, s) \|$ converges to 0 in probability. This then implies by the triangle inequality that for any sequence $\theta_T \in \Theta$
\[
\| G_T(\theta_T, P_k, s^N) \| - \| G_T(\theta_T, P_0, s) \| \leq \| G_T(\theta_T, P_k, s^N) - G_T(\theta_T, P_0, s) \| + o_p(1) \leq \| G_T(\theta_T, P_k, s^N) - G_T(\theta_T, P_0, s) \| + o_p(1).
\]

Let $\hat{\theta} = \arg \min_{\theta \in \Theta} \| G_T(\theta, P_0, s) \|$. Then $\hat{\theta}$ satisfies $\| G_T(\hat{\theta}, P_0, s) \| = \inf_{\theta \in \Theta} \| G_T(\theta, P_0, s) \| + o_p(1)$ because
\[
0 \leq \| G_T(\hat{\theta}, P_0, s) \| - \inf_{\theta \in \Theta} \| G_T(\theta, P_0, s) \| = \| G_T(\hat{\theta}, P_0, s) \| - \| G_T(\hat{\theta}, P_0, s) \| + o_p(1) \leq \| G_T(\hat{\theta}, P_0, s^N) \| - \| G_T(\hat{\theta}, P_0, s) \| + o_p(1) = o_p(1).
\]

Hence, by the first part, proving $\sup_{\delta \in \Theta} \| G_T(\theta_T, P_k, s^N) - G_T(\theta_T, P_0, s) \| = o_p(1)$ is sufficient for consistency. Let $Z$ be the $\sum_{t=1}^T K \times d_g$ matrix of instruments. By the Cauchy–Schwarz inequality,
\[
\| G_T(\theta_T, P_k, s^N) - G_T(\theta_T, P_0, s^N) \|^2
\leq \frac{1}{T} \| Z(\xi(\theta, P_k, s^N) - \xi(\theta, P_0, s^N)) \|^2
\leq \frac{1}{T} \| Z \| \times \frac{1}{T} \| \xi(\theta, P_k, s^N) - \xi(\theta, P_0, s^N) \|^2.
\]

Since $\frac{1}{T} \| Z \| = o_p(1)$ by Assumption A.5, it suffices to prove that
\[
\sup_{\theta \in \Theta} \| \xi(\theta, P_k, s^N) - \xi(\theta, P_0, s) \|^2
\leq \sup_{\theta \in \Theta} \frac{1}{T} \sum_{t=1}^T \| \xi_T(\theta, P_k, s^N) - \xi_T(\theta, P_0, s) \|^2 = o_p(1).
\]

By Assumption A.7 we have
\[
\sup_{\theta \in \Theta} \| \xi_T(\theta, P_k, s^N) - \xi_T(\theta, P_0, s) \|^2
\leq \sup_{\theta \in \Theta} \| \xi_T(\theta, P_k, s^N) - \xi_T(\theta, P_0, s) \|^2 = o_p(1).
\]

This then implies that $\sup_{\theta \in \Theta} \| \xi_T(\theta, P_k, s^N) - \xi_T(\theta, P_0, s) \|^2 = o_p(1)$ because by Assumption A.8, if instead $\max_{1 \leq t \leq T} \| \xi_T(\theta, P_k, s^N) - \xi_T(\theta, P_0, s) \| > \delta$, then
\[
\sup_{\theta \in \Theta} \| \xi_T(\theta, P_k, s^N) - \xi_T(\theta, P_0, s) \| > \delta,
\]

with probability approaching 1, which contradicts (11). \hfill \Box

B.2. Asymptotic normality

Next I present sufficient conditions for the asymptotic normality results. To do so define
\[
\varepsilon_{\text{ort}} \equiv v_T(\theta_0, \xi_T(\theta_0, P_0, s_t), \nu_T)
\]
and
\[
\varepsilon_{\text{ort}} \equiv \frac{\partial v_T(\theta_0, \xi_T(\theta_0, P_0, s_t), \nu_T)}{\partial \xi} dP_0(v).
\]
Also let $e_j$ denote the jth column of the $J_t \times J_t$ identity matrix and define

$$H_{i\theta} = \frac{\partial \sigma_j(\theta_0, \xi_j(\theta_0, P_{\theta_0}), P_{\theta_0})}{\partial \xi_j}.$$ 

$$I_{0\theta} = \sum_{i=1}^{j} H_{i\theta}^{-1} K_{0\theta} H_{i\theta}^{-1} e_j e_j' H_{i\theta}^{-1},$$

and $K_{0\theta} = \frac{\partial^2 \sigma_j(\theta_0, \xi_j(\theta_0, P_{\theta_0}, s_i), P_{\theta_0})}{\partial \xi_j \partial \xi_k}.$

For any function $h(v, x)$ denote by $E_i^r(h(v, x))$ the expectation with respect to $P_{\theta_0}$, given the data.

**Assumption B.1.** $\theta_0$ is an interior point of $\Theta$.

**Assumption B.2.** For all $P \in \mathcal{P}$, the function $G_r(\theta, P, s)$ is differentiable at $\theta_0$. Define the derivative matrix $\Gamma_t = \frac{\partial G_r(\theta_0, P_{\theta_0})}{\partial \theta_0}$. If $T \sum_{t=1}^{T-1} \Gamma_t$ converges to a matrix, $\Gamma$, of full rank.

**Assumption B.3.** $v_t (\theta, \xi_t, v)$ is four times continuously differentiable in $\xi_t$ for all $\xi_t \in \mathbb{R}^d, \theta \in \Theta$, and $v \in \mathbb{R}^d$, $v_t (\theta, \xi_t, v)$ and its first four derivatives with respect to $\xi_t$ are bounded, continuous, and differentiable with respect to $\theta$ for all $\xi_t \in \mathbb{R}^d, \theta \in \Theta$, and $v \in \mathbb{R}^d$.

**Assumption B.4.** $s_{jm} \in \{0, 1\}$ are iid draws with $E(s_{jm}) = s_T$, the draws are independent of $(p_t, z_t, \xi_t)$ conditional on $s_t$, and $\ln(T)/N \rightarrow 0$.

**Assumption B.5.** As $T \rightarrow \infty$

$$\sup_{\theta \in \Theta} \max_{1 \leq t \leq T, 1 \leq j \leq k} \left| \frac{\partial \sigma_j(\theta, \xi_j(\theta, P_{\theta_0}, s_t), P_{\theta_0})}{\partial \xi_j} \right| \rightarrow 0,$$

$$\sup_{\theta \in \Theta} \max_{1 \leq t \leq T, 1 \leq j \leq k} \left| \frac{\partial^2 \sigma_j(\theta, \xi_j(\theta, P_{\theta_0}, s_t), P_{\theta_0})}{\partial \xi_j \partial \xi_j} \right| \rightarrow 0,$$

$$\sup_{\theta \in \Theta} \max_{1 \leq t \leq T, 1 \leq j \leq k, l \leq m} \left| \frac{\partial^3 \sigma_j(\theta, \xi_j(\theta, P_{\theta_0}, s_t), P_{\theta_0})}{\partial \xi_j \partial \xi_l \partial \xi_m} \right| \rightarrow 0.$$

**Assumption B.6.** $v_t$ is iid across $r$, independent across $t$, and independent of $(p_t, z_t, \xi_t)$.

**Assumption B.7.** The random variables in $x_t$ have bounded support and $E(x_{jT}^2) \leq M$ for some constant $M$ and for all $l = 1, \ldots, d^2$, $t = 1, \ldots, T$ and $j = 1, \ldots, J_t$.

**Assumption B.8.** The absolute value of each element of $H^{-1}$ is bounded above by some constant $M$ in a neighborhood of $\theta_0$ for all $t = 1, \ldots, T$.

**Assumption B.9.** There exists a $J_t \times d_0$ matrix $H_{1t}(v)$ and a $d_0 \times d_0$ matrix $H_{2t}(v)$ such that for all $j = 1, \ldots, J_t$,

$$\left| \frac{\partial v_j(\theta, \xi_j, v)}{\partial \theta} \right| \leq H_{1t}(v),$$

$$\left| \frac{\partial^2 v_j(\theta, \xi_j, v)}{\partial \theta \partial \theta^\prime} \right| \leq H_{2t}(v),$$

and each element of $H_{1t}(v)$ and $H_{2t}(v)$ has 4 bounded absolute moments with respect to $P_{\theta_0}$. The inequalities are understood element by element and the elements of $H_{1t}(v)$ do not depend on $t$.

**Assumption B.10.** Assume that there are positive definite matrices $\Phi_1$, $\Phi_2$, and $\Phi_3$ such that

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} E \left( \frac{1}{T} \sum_{t=1}^{T} z_t^\prime \xi_t(\theta_0, P_{\theta_0}, s_t) \xi_t(\theta_0, P_{\theta_0}, s_t) z_t \right) = \Phi_1,$$

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} \text{Var} \left( z_t^\prime H_{1t}^{-1} e_t \right) = \Phi_2,$$

and

$$\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} \text{Var} \left( z_t^\prime H_{1t}^{-1} (s_t - s_t) \right) = \Phi_3.$$

**Assumption B.11.** Assume that the following limit exists

$$\bar{\mu}_1 = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{n_j} E \left( z_j^\prime \left( \frac{1}{E} H_{1j}^{-1} E_i \left( d_{0\theta} H_{1j}^{-1} e_t \right) \right) \right),$$

and

$$\bar{\mu}_2 = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{j=1}^{n_j} E \left( z_j^\prime \left( \text{diag}(I_{0\theta}) s_t - s_t I_{0\theta} s_t \right) \right).$$

The first two assumptions are standard. **Assumption B.3** assumes smoothness of the function generating the market shares and can easily be checked once more structure is imposed. It rules out the pure characteristics model, where $v_{jt}$ is an indicator function. **Assumption B.4** states that the observed shares are generated from a random sample of $N$ consumers. **Assumption B.5** is very similar to **Assumption A.7**. **Assumption B.6** states that the draws, $v_{jt}$, are iid and differ across markets. **Assumption B.7** states that the prices and product characteristics have bounded support and that the instruments have four bounded moments. **Assumption B.8** states that the elements of the inverse of a matrix are bounded. This assumption generally holds under the assumptions in this paper if the inverse exists. **Assumption B.9** is easily verified once more structure on the function generating the market shares is imposed. **Assumptions B.10** and **B.11** place restrictions on the data generating process. They guarantee the applicability of central limit theorems and the convergence of the bias terms in the asymptotic expansion, respectively. Both assumptions are implied by the other assumptions if the distributions of the data and the random coefficients are the same in all markets.

These assumptions lead to the main theorem of this section.

**Theorem A.2.** Assume that **Assumptions A.1–A.9** and **B.1–B.11** hold. Then

$$\sqrt{T} \left( \hat{\theta} - \theta_0 \right) = \left( \Gamma' W \Gamma \right)^{-1} \Gamma' W + o_p(1)$$

$$\times \left( Q_{1T} + \frac{1}{\sqrt{R}} Q_{2T} + \frac{1}{\sqrt{N}} Q_{3T} + \sqrt{\frac{r}{N}} C_{1T} + \sqrt{\frac{r}{N}} C_{2T} \right),$$

where $Q_{1T} \xrightarrow{p} \bar{\mu}_1$, and $C_{1T} \xrightarrow{p} \bar{\mu}_2$. Furthermore, $Q_{1T}, Q_{2T},$ and $Q_{3T}$ are asymptotically independent.
The proof now consists of two parts. First I show that
\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) \xrightarrow{d} N \left( \left( \Gamma'W\Gamma \right)^{-1} \Gamma'W \left( \lambda_1 \hat{\mu}_1 + \lambda_2 \hat{\mu}_2 \right), V_1 \right)
\]
where
\[
V_1 = \left( \Gamma'W\Gamma \right)^{-1} \Gamma'W\Phi_W \Gamma' \left( \Gamma'W\Gamma \right)^{-1}.
\]

**Proof of Theorem A.2.** The objective is to minimize \( G_T(\theta, P_R, s^N) \) \( W_I G_T(\theta, P_R, s^N) \). The first order condition is
\[
\left( \frac{\partial}{\partial \theta} G_T(\theta, P_R, s^N) \right)' W_I G_T(\hat{\theta}, P_R, s^N) = 0.
\]

Define
\[
D_T(\theta, P_R, s^N) \equiv \frac{\partial}{\partial \theta} G_T(\theta, P_R, s^N).
\]
The mean value theorem implies that a first order expansion of \( D_T(\hat{\theta}, P_R, s^N) \) around \( \theta = \theta_0 \) yields
\[
D_T(\hat{\theta}, P_R, s^N) W_I \left( G_T(\theta_0, P_R, s^N) + D_T(\hat{\theta}, P_R, s^N) \left( \hat{\theta} - \theta_0 \right) \right) = 0
\]
where \( \hat{\theta} \) is between \( \theta_0 \) and \( \hat{\theta} \). Thus
\[
\sqrt{T} \left( \hat{\theta} - \theta_0 \right) = \left( D_T(\hat{\theta}, P_R, s^N) W_I D_T(\hat{\theta}, P_R, s^N) \right)^{-1} \times \left( D_T(\hat{\theta}, P_R, s^N) W_I \sqrt{T} G_T(\theta_0, P_R, s^N) \right).
\]
The proof now consists of two parts. First I show that
\[
\sqrt{T} G_T(\theta_0, P_R, s^N) = \sqrt{T} G_T(\theta_0, P_R, s) + O_p \left( \frac{1}{\sqrt{T}} \right) + O_p \left( \frac{1}{N} \right) + O_p \left( \frac{\sqrt{T}}{R} \right) + O_p \left( \frac{\sqrt{T}}{N} \right)
\]
and I derive an expressions for the terms on the right hand side. Next I prove that for any consistent estimator \( \hat{\theta} \) of \( \theta \) it holds that \( \hat{\theta}_T(\hat{\theta}, P_R, s^N) \) converges to \( \Gamma \) in probability. Combining these results yields the conclusion of the theorem.

We have
\[
\sqrt{T} G_T(\theta_0, P_R, s^N) = \sqrt{T} G_T(\theta_0, P_R, s) + \sqrt{T} G_T(\theta_0, P_R, s) - G_T(\theta_0, P_R, s).
\]
The first term is \( O_p(1) \) and belongs to the GMM objective function without simulation error. By Assumptions A.4 and B.7–B.10 and Lyapunov’s CLT it converges to a normally distributed random variable. Next consider
\[
G_T(\theta_0, P_R, s) = -G_T(\theta_0, P_R, s)
\]
where
\[
= \frac{1}{T} \sum_{i=1}^{T} \sum_{j=1}^{I} z_{ij} (\xi_{i\theta} (\theta_0, P_{R_i}, s_i) - \xi_{i\theta} (\theta_0, P_{R_i}, s_i))
\]

16 The notation is a bit simplified because \( \hat{\theta} \) differs for each element of \( D_I(\theta, P_R, s^N) \). See Lemma A.4.
17 Since \( \left( \begin{array}{c}
\frac{\partial}{\partial \theta_0} G_T(\theta_0, P_R, s^N)
\end{array} \right) \) is positive definite, Assumptions B.7–B.9 guarantee the existence of 2 + \( \delta \) moments, which is needed for Lyapunov’s CLT to apply.
\((\theta_0, P_{0t}, s_t), P_{0s}\) the previous expansion implies that there exists \(\epsilon_t \in (0, 1)\) such that

\[
\xi_t(\theta_0, P_{0t}, s_t) = \xi_t(\theta_0, P_{0t}, s_t) - e_{0t}'H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t} + \frac{1}{2}e_{0t}'H_{0t}^{-1}\left(\frac{\partial^2 \xi_t}{\partial \xi^2} + \frac{\partial^2 \xi_t}{\partial \xi \partial \xi}\right)H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t}
\]

where \(e_{0t}\) is the \(j\)th column of the \(J_t \times J_t\) identity matrix,

\[
H_{0t} = \frac{\partial \xi_t}{\partial \xi} \quad \text{and} \quad H_{0t} = \frac{\partial \xi_t}{\partial \xi},
\]

error\(_{0t} = e_{0t}'(H_{0t}^{-1} - H_{0t}^{-1})(H_{0t}^{-1} - H_{0t})H_{0t}^{-1}e_{0t} + \frac{1}{2}e_{0t}'H_{0t}^{-1}\left(\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t}
\]

\[
\times \frac{1}{2}e_{0t}'H_{0t}^{-1}\left(\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t}
\]

\[
+ \sum_{\alpha(t) \leq |t|} \alpha! \frac{\partial^{|t|} \sigma_t^{-1}(\theta_0, s_t)}{\partial \xi^{|t|}} \left|_{\xi = \xi_0(t) \theta_0, P_{0t}, s_t)}p_{0t} e_{0t}'H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t}
\]

Now define

\[
K_{0t} = \left(\frac{\partial^2 \sigma_t}{\partial \xi^2} \frac{\partial \xi_t}{\partial \xi}\right) \quad \text{and} \quad I_{0t} = \sum_{t=1}^h H_{0t}^{-1}K_{0t}H_{0t}^{-1}e_{0t}e_{0t}'H_{0t}^{-1}.
\]

Then

\[
\xi_t(\theta_0, P_{0t}, s_t) = \xi_t(\theta_0, P_{0t}, s_t) - e_{0t}'H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t} + \frac{1}{2}e_{0t}'H_{0t}^{-1}\left(\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t}
\]

It now follows that \(\sqrt{T}(G_t(\theta_0, P_{0t}, s) - G_t(\theta_0, P_{0t}, s))\) equals

\[
- \frac{1}{\sqrt{T}} \sum_{t=1}^T z_{it}'H_{0t}^{-1}e_{0t} + \frac{1}{\sqrt{T}} \sum_{t=1}^T \sum_{h=1}^h z_{ih}'
\]

\[
\times \left(e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t} - \frac{1}{2}e_{0t}'e_{0t}\right) + \frac{1}{\sqrt{T}} \sum_{t=1}^T z_{it}'e_{0t}.
\]

Now we have three terms. The first term is \(O_p\left(\frac{1}{\sqrt{T}}\right)\) and converges to a normally distributed random variable. The second term is \(O_p\left(\frac{1}{\sqrt{T}}\right)\) and converges to a normally distributed random variable. The third term is \(O_p\left(\frac{1}{T}\right)\) and converges to a normally distributed random variable. By Assumption B.11

\[
\lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^T \sum_{j=1}^h E_z\left(\xi_{0t}'H_{0t}^{-1}e_{0t} + e_{0t}'H_{0t}^{-1}\frac{\partial e_{0t}}{\partial \xi}H_{0t}^{-1}e_{0t}
\]

\[
- \frac{1}{2}E_z(e_{0t}'e_{0t})\right) = \mu_1.
\]

By Assumptions B.6–B.8 and B.11, Chebyshev’s law of large numbers (LLN), and Lyapunov’s CLT

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^T \sum_{j=1}^h z_{jt}'e_{0t}e_{0t} \to \mu_1,
\]

which implies that the second term times \(\frac{1}{\sqrt{T}}\) converges in probability to a constant.

Next I will prove that

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^T \sum_{j=1}^h z_{jt}'e_{0t}e_{0t} = o_p\left(\frac{\sqrt{T}}{T}\right).
\]

Recall that

\[
= \frac{1}{\sqrt{T}} \sum_{t=1}^T \sum_{j=1}^h z_{jt}'e_{0t}e_{0t} = o_p\left(\frac{\sqrt{T}}{T}\right).
\]

By Assumptions B.5 and B.8, \(H_{rt}^{-1} = H_{or}^{-1} + o_p(1)\) where each element of the \(J_t \times J_t\) term does not depend on \(t\). Similarly

\[
\sum_{t=1}^T \sum_{j=1}^h z_{jt}'e_{0t}e_{0t} = o_p\left(\frac{\sqrt{T}}{T}\right).
\]

where each element of the \(J_t \times J_t\) term does not depend on \(t\). It now follows from Assumptions B.6–B.8 as well as Lemma A.3 that

\[
\frac{1}{\sqrt{T}} \sum_{t=1}^T \sum_{j=1}^h z_{jt}'e_{0t}e_{0t} \to \mu_1.
\]
and
\[
1 + 2 - 1 \frac{T}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{h} \frac{\sigma'_R}{\sigma_R} \left( H^{-1}_{Rt} \times \left( \sum_{k=1}^{h} \left( H^{-1}_{Rt} \left( \frac{\partial^2 \sigma_{Rt}}{\partial \xi_{t} \partial \xi_{k}} \left( \theta_0, \xi_{t}(\theta_0, P_{Rt}, s_t), P_{Rt} \right) \right) \right) \right) \right) \right) e_{Rt} = o_p \left( \sqrt{\frac{1}{R}} \right).
\]

Finally consider
\[
1 + 2 - 1 \frac{T}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{h} \frac{1}{\alpha_t^2} \left( \frac{\partial |\sigma_{Rt}^{-1}(\theta_0, s, P_{Rt})|}{\partial s_{Rt}} \right) \right) \right) e_{Rt}.
\]

Similarly, we can write the third partial derivatives of the inverse function as a function of partial derivatives of \( \sigma_t(\theta_0, \xi, P_{Rt}) \) evaluated at \( \tilde{\xi}_t \), which satisfies
\[
\sigma_R(\theta_0, \xi_0, P_{Rt}) = \sigma_R(\theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt}) + c_{\xi} (\xi_{t} - \sigma_R(\theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt})).
\]

Since \( \max_{1 \leq t \leq T} \left| \sigma_t - \sigma_R(\theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt}) \right| = o_p(1) \) also \( \max_{1 \leq t \leq T} \left| \sigma_t - \sigma_R(\theta_0, \xi_0, P_{Rt}) \right| = o_p(1) \), which is equivalent to
\[
\max_{1 \leq t \leq T} \left| \sigma_t(\theta_0, \xi_0, P_{Rt}) - \sigma_R(\theta_0, \xi_0, P_{Rt}) \right| = o_p(1).
\]

This implies by Assumption A.8 that \( \max_{1 \leq t \leq T} \left| \frac{\delta^3 \sigma_t}{\delta s^3}_{Rt}(\theta_0, \xi_0, P_{Rt}) - \frac{\delta^3 \sigma_{Rt}}{\delta s^3}_{Rt}(\theta_0, \xi_0, P_{Rt}) \right| = o_p(1) \), which in turn means that \( \max_{1 \leq t \leq T} \left| \frac{\delta^3 \sigma_{Rt}}{\delta s^3}_{Rt}(\theta_0, \xi_t, P_{Rt}) - \frac{\delta^3 \sigma_{Rt}}{\delta s^3}_{Rt}(\theta_0, \xi_t, P_{Rt}) \right| = o_p(1) \). Next, it is easy to verify that
\[
\frac{\partial |\sigma_{Rt}^{-1}(\theta_0, s, P_{Rt})|}{\partial s_{Rt}} \right) \right) e_{Rt}.
\]

is a function of \( \left( \frac{\partial |\sigma_{Rt}^{-1}(\theta_0, s, P_{Rt})|}{\partial s_{Rt}} \right)_{\xi_t=\xi_t}^{-1} \) for all \( |\alpha| \) as partial derivatives of \( \sigma_t(\theta_0, \xi, P_{Rt}) \) up to order 3 evaluated at \( \tilde{\xi}_t \). But since \( \max_{1 \leq t \leq T} \left| \xi_t - \xi_t(\theta_0, P_{Rt}, s_t) \right| = o_p(1) \), it follows from Assumptions B.3 and B.5 that
\[
\frac{\partial |\sigma_{Rt}^{-1}(\theta_0, s, P_{Rt})|}{\partial s_{Rt}} \right) \right) e_{Rt} = Q_0(0,0) + o_p(1)
\]

where \( Q_0(0,0) \) is bounded for each \( t \) and \( |\alpha| \) and the \( o_p(1) \) term does not depend on \( t \). As a consequence, it follows from Lemma A.3 that
\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{h} \frac{1}{\alpha_t^2} \left( \frac{\partial |\sigma_{Rt}^{-1}(\theta_0, s, P_{Rt})|}{\partial s_{Rt}} \right) \right) e_{Rt} = o_p \left( \sqrt{T} \right).
\]

Next consider the last term of (12), namely
\[
\sqrt{T} \left( G_{T}(\theta_0, P_{R}, s_N) - G_{T}(\theta_0, P_{R}, s) \right)
\]
\[
= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{i=1}^{h} \left( \frac{\partial^3 \sigma_{Rt}}{\partial \xi_{t} \partial \xi_{k}} \left( \theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt} \right) \right) \left( H^{-1}_{Rt} e_{Rt} e_{Rt}' H^{-1}_{Rt} \right) \right) \right) \right) e_{Rt} = o_p \left( \sqrt{T} \right).
\]

Finally consider
\[
\sqrt{T} \left( \tilde{H}_{Rt} - H_{Rt} \right)
\]
\[
= \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{h} \left( \frac{\partial^3 \sigma_{Rt}}{\partial \xi_{t} \partial \xi_{k}} \left( \theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt} \right) \right) \left( H^{-1}_{Rt} e_{Rt} e_{Rt}' H^{-1}_{Rt} \right) \right) \right) \right) e_{Rt} = o_p \left( \sqrt{T} \right).
\]

Similarly, the third terms do not depend on \( t \) or \( j \). Moreover
\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{h} \left( \frac{\partial^3 \sigma_{Rt}}{\partial \xi_{t} \partial \xi_{k}} \left( \theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt} \right) \right) \left( H^{-1}_{Rt} e_{Rt} e_{Rt}' H^{-1}_{Rt} \right) \right) \right) \right) e_{Rt} = o_p \left( \sqrt{T} \right).
\]

Also, just as before,
\[
\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \sum_{j=1}^{h} \left( \frac{\partial^3 \sigma_{Rt}}{\partial \xi_{t} \partial \xi_{k}} \left( \theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt} \right) \right) \left( H^{-1}_{Rt} e_{Rt} e_{Rt}' H^{-1}_{Rt} \right) \right) \right) \right) e_{Rt} = o_p \left( \sqrt{T} \right).
\]

Finally consider
\[
D_T(\hat{\theta}, P_{R}, s_N) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( \frac{\partial^3 \sigma_{Rt}}{\partial \xi_{t} \partial \xi_{k}} \left( \theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt} \right) \right) \left( H^{-1}_{Rt} e_{Rt} e_{Rt}' H^{-1}_{Rt} \right) \right) \right) e_{Rt} \right) \right) e_{Rt} = o_p \left( \sqrt{T} \right).
\]

To show that for any consistent estimator, \( \hat{\theta} \), of \( \theta \), \( D_T(\hat{\theta}, P_{R}, s_N) \) \( P \rightarrow \Gamma \) it is sufficient to show (by Assumptions B.2 and B.9) that
\[
D_T(\theta, P_{R}, s_N) = \frac{1}{\sqrt{T}} \sum_{t=1}^{T} \left( \frac{\partial^3 \sigma_{Rt}}{\partial \xi_{t} \partial \xi_{k}} \left( \theta_0, \xi(\theta_0, P_{Rt}, s_t), P_{Rt} \right) \right) \left( H^{-1}_{Rt} e_{Rt} e_{Rt}' H^{-1}_{Rt} \right) \right) \right) e_{Rt} \rightarrow \Gamma \text{ uniformly over } \theta \text{ in a neighborhood of } \theta_0. \]

This then implies that
\[
D_T(\hat{\theta}, P_{R}, s_N)' W_T D_T(\hat{\theta}, P_{R}, s_N) \rightarrow \Gamma \text{ uniformly over } \theta \text{ in a neighborhood of } \theta_0. \]

Finally
\[
D_T(\hat{\theta}, P_{R}, s_N)' W_T D_T(\hat{\theta}, P_{R}, s_N) \rightarrow \Gamma \text{ uniformly over } \theta \text{ in a neighborhood of } \theta_0. \]
Notice that
\[
D_T(\theta, P_R, s^N) - \frac{1}{T} \sum_{t=1}^{T} \frac{\partial G_T(\theta, P_{0t})}{\partial \theta} \\
= D_T(\theta, P_R, s^N) - \frac{\partial G_T(\theta, P_{0t}, s)}{\partial \theta} \\
+ \frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial \xi_t(\theta, P_{0t}, s_t)}{\partial \theta} \right) - E \left( \frac{\partial \xi_t(\theta, P_{0t}, s_t)}{\partial \theta} \right)
\]
and that
\[
\frac{1}{T} \sum_{t=1}^{T} \frac{\partial \xi_t(\theta, P_{0t}, s_t)}{\partial \theta} - E \left( \frac{\partial \xi_t(\theta, P_{0t}, s_t)}{\partial \theta} \right) \overset{p}{\to} 0
\]
uniformly over \( \theta \) by Assumptions A.4, B.8 and B.9 and the ULLN of Andrews (1987).\(^{18}\)

Next rewrite
\[
D_T(\theta, P_R, s^N) = \frac{\partial G_T(\theta, P_{0t}, s)}{\partial \theta} \\
= \frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial \xi_t(\theta, P_{0t}, s_t)}{\partial \theta} \right) \\
- \frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} \right) \\
\times \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} \\
- \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta}
\]
where \( \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} \) is the derivatives with respect to the first element only. Similar as before
\[
\frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \xi_t} = \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} + o_p(1)
\]
in a neighborhood of \( \theta_0 \), where the \( o_p(1) \) term does not depend \( \theta \) or \( t \). Now by Assumption B.8
\[
\left( \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \xi_t} \right)^{-1} \\
= \left( \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} \right)^{-1} + o_p(1),
\]
and the \( o_p(1) \) term does not depend \( \theta \) or \( t \). By Hölder’s inequality and Assumptions A.4, B.7 and B.9 is \( o_p(1) \). Hence,
\[
D_T(\theta, P_{0t}) = \frac{\partial G_T(\theta, P_{0t}, s)}{\partial \theta} \\
= -\frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} \right)^{-1} \\
\times \left( \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} \right) \\
+ \frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial \sigma_t\left(\theta, \xi_t(\theta, P_{0t}, s_t), P_{0t}\right)}{\partial \theta} \right)^{-1}
\]
It follows from Assumptions A.7, A.8 and B.9 that for all \( j \)
\[
\left| \frac{\partial \sigma_{Rj}(\theta, \xi^N, P_{0t}, s)}{\partial \theta} \right| \leq \frac{1}{R} \sum_{r=1}^{R} \left| \xi_t(\theta, P_{0t}, s^N) - \xi_t(\theta, P_{0t}, s_t) \right| = o_p(1)
\]
uniformly over \( \theta \) and \( t \). Thus, by Hölder’s inequality and Assumptions A.4, B.7 and B.9
\[
\frac{1}{T} \sum_{t=1}^{T} \left( \frac{\partial \sigma_t(\theta, \xi_t, P_{0t}, s_t), P_{0t})}{\partial \xi_t} \right)^{-1} \\
\times \left( \frac{\partial \sigma_t(\theta, \xi_t, P_{0t}, s_t), P_{0t})}{\partial \theta} \right) \\
= \frac{1}{T} \sum_{t=1}^{T} \sum_{r=1}^{R} \left( \frac{\partial \sigma_t(\theta, \xi^N, P_{0t}, s^N), P_{0t})}{\partial \xi_t} \right)^{-1} \\
\times \left( \frac{\partial \sigma_t(\theta, \xi_t, P_{0t}, s_t), P_{0t})}{\partial \theta} \right) \\
- \frac{1}{T} \sum_{t=1}^{T} \frac{1}{R} \sum_{r=1}^{R} \left( \frac{\partial \sigma_t(\theta, \xi^N, P_{0t}, s^N), P_{0t})}{\partial \xi_t} \right)^{-1} \\
\times \left( \frac{\partial \sigma_t(\theta, \xi_t, P_{0t}, s_t), P_{0t})}{\partial \theta} \right)
\]
with \( o_p(1) \) uniformly over \( \theta \) by the uniform law of large numbers of Andrews (1987) and Assumptions A.4 and B.6–B.9. Hence \( \mathrm{sup}_{P_0 \in \mathcal{P}} \left| D_T(\theta, P_R) - \frac{\partial G_T(\theta, P_{0t})}{\partial \theta} \right| = o_p(1) \). \( \Box \)

B.3. Bias correction

Theorem 2 also holds in the general case under Assumptions A.1–A.9 and B.1–B.11, where expressions for \( \hat{\mu}_1 \) and \( \hat{\mu}_2 \) are given in Section 3. In the last part of the proof of Theorem 2, I show that for any consistent estimator, \( \hat{\theta} \), of \( \theta \), it holds that \( \frac{1}{T} \sum_{t=1}^{T} \frac{\partial G_T(\theta, P_{0t}, s_t)}{\partial \theta} \rightarrow^D \mathcal{G} \). Analogous arguments can be used here to show \( \hat{\mu}_1 \rightarrow^D \mu_1 \) as \( T \rightarrow \infty \) and \( \hat{\mu}_2 \rightarrow^D \mu_2 \) as \( T \rightarrow \infty \).

Appendix C. Proofs of Theorems 1 and 2

I will first prove that the assumptions imply that \( \xi_t(\theta, P_{0t}, s_t, x_t) \) is an element of a bounded subset of \( \mathbb{R}^k \) and then use this result to verify the assumptions of Theorem A.2. First decompose \( w_k = \left( w_{k1}, w_{k2} \right) \) where \( w_{k1} \sim Q \) has compact support and \( w_{k2} \sim F \) with \( F \) in \( \mathbb{P} \times \cdots \times \mathbb{P} \). Let \( \mathcal{X} \) be the support of \( x_t \). Now assume that \( \xi_t(\theta, P_{0t}, s_t, x_t) \) is not an element of a bounded subset of \( \mathbb{R}^k \). Then, there exists a sequence \( \{ s^k, x^k, \theta^k \} \), \( k = 1, 2, \ldots, \) with \( \epsilon \leq s^k \leq 1 - \epsilon, j = 0, \ldots, J, \theta \in \Theta, x^k \in \mathcal{X} \) such that for some \( j \),
\[
\xi^k_t \equiv \xi(\theta^k, P_{0t}, s^k, x^k) \rightarrow \infty \text{ or } \xi^k_t \rightarrow -\infty
\]
and
\[
s^k = \int_{\mathcal{P}} \phi_k(\theta^k, x^k, \xi^k, g(a^k, w_1, w_2)) dQ_{0t}(w_1) dF(w_2).
\]
Next define
\[
\xi_j^k \equiv \sup_{\theta \in \Theta} \sup_{(x, \xi) \in \mathbb{X}} \sup_{u \in \supp(Q)} \sup_{w \in A} \int \nu_{(x, \xi, \theta, \gamma(a, (u_1, u_2)))} \exp(dF(w_2)).
\]
If \( \xi_j^k \to \infty \), then \( s^0 k_0 \leq s^0(\xi_j^k) \to 0 \), which is a contradiction. Similarly if \( \xi_j^k \to -\infty \), then \( s_j^k \to 0 \).

I now verify the assumptions of Theorem A.2. Assumption A.1 holds in the BLP model as shown by Berry (1994). The result also holds with approximated predicted shares and observed shares, as long as the observed shares are positive with probability one, which is guaranteed by Assumption RC4. Assumptions A.2 and A.3 clearly hold. Assumption A.4 is assumed in Assumption RC1. Assumption A.5 is implied by Assumption RC7 and Assumption A.6 holds because \( \hat{\xi}(\theta_0, P_{\theta_0}, s_t) \) is bounded. The first part of Assumption A.7 follows from Lemma A.1 (where \( x_t = (x_t, \xi_t(\theta_0, P_{\theta_0}, s_t, x_t)) \) and \( f(x, \theta, v) = v_1(\theta, x, \xi, v) \)). The conditions of the lemma are implied by Assumptions RC2–RC6.

The second part holds because by the Bernstein inequalities there exists a constant \( c \) such that
\[
\Pr \left( \max_j \sum_{t \in T} \left| s_R^t - \hat{s}_t \right| > \varepsilon \right) \leq 2 \sum_{t \in T} \sum_{j=1}^p \Pr \left( \left| s_R^t - s_t \right| > \varepsilon \right) 
\leq 2 \sum_{t \in T} \exp(-cN).
\]

To see why Assumption A.8 holds, take \( \xi_t \) such that \( \|\xi_t - \hat{\xi}_t(\theta, P_{\theta_0}, s_t^k)\| \geq \beta \). Assume without loss of generality that \( \|\xi_t - \hat{\xi}_t(\theta, P_{\theta_0}, s_t^k)\| \leq \|\xi_t - \hat{\xi}_t(\theta, P_{\theta_0}, s_t)\| \geq \|\hat{\xi}_t - \xi_t(\theta, P_{\theta_0}, s_t^k)\| \geq \beta \) but similar arguments apply to \( \hat{\xi}_t - \xi_t(\theta, P_{\theta_0}, s_t) \leq -\beta \). Let \( \xi_t + \beta = (\hat{\xi}_t + \beta, \ldots, \hat{\xi}_t + \beta) \) and notice that for all \( \beta > 0 \)
\[
\exp(\beta) \int v_1(\theta, x_t, \xi_t, v) \exp(dP_0) = \int v_1(\theta, x_t, \xi_t + \beta, v) \exp(dP_0) > \int v_1(\theta, x_t, \xi_t, v) \exp(dP_0)
\]
and
\[
\int v_1(\theta, x_t, \xi_t, v) \exp(dP_0) = \int v_1(\theta, x_t, \xi_t + \beta, v) \exp(dP_0) > \int v_1(\theta, x_t, \xi_t, v) \exp(dP_0).
\]

Also notice that by Assumptions RC2 and RC4–RC6 there exists \( \eta > 0 \) and a compact set \( V_1 \subset \mathbb{R}^d \), such that with probability approaching 1, \( \eta < v_1(\theta, x_t, \xi_t, (\theta, P_{\theta_0}, s_t^k, v), v) < 1 - \eta \) for all \( j \) and \( v \in V_1 \), and \( \int v_1(\theta, x_t, \xi_t, v) \exp(dP_0) > p > 0 \).

Next define \( \delta_0 = \min \left\{ \delta, \frac{1}{\eta} \ln \left( \frac{1}{\eta^2} \right) \right\} \) and \( C(\delta) = p\eta(1 - \exp(2\delta_0)(1 - \eta)) \delta_0 > 0 \). The previous arguments imply that for some \( \delta \in (0, \delta_0) \) it holds that with probability approaching 1
\[
\int v_1(\theta, x_t, \xi_t, v) \exp(dP_0) > \int v_1(\theta, x_t, \xi_t + \beta, v) \exp(dP_0) \geq \int v_1(\theta, x_t, \xi_t, v) \exp(dP_0) > \int v_1(\theta, x_t, \xi_t, v) \exp(dP_0).
\]
This proves that Assumption A.9 holds. Assumption A.10 is assumed in Assumption RC9.

Assumption B.1 is directly assumed. Differentiability in Assumption B.2 holds. Full rank is assumed in Assumption RC9 and the limit result holds by Assumption RC8 (the moment conditions in Assumptions RC5–RC7 guarantee that the moment condition in Assumption RC8 holds). Assumption B.3 holds since the derivatives are bounded. Assumptions B.4, B.6 and B.7 are easily interpretable and directly assumed. Assumption B.5 holds by Lemma A.1, just like Assumption A.7. By the contraction mapping property of the BLP model, \( \frac{\delta_0}{\eta^2} \frac{\delta_0}{\eta^2}(\theta, P_{\theta_0}, s_t^k) \) is invertible (see Dube et al., 2012). Thus, the determinant is not equal to 0 for all \( \theta, s_t, x_t \). By Assumptions RC5 and RC6 and continuity of the determinant in all arguments, it follows that the determinant is bounded away from 0. Since the elements of \( \frac{\delta_0}{\eta^2} \frac{\delta_0}{\eta^2}(\theta, P_{\theta_0}, s_t^k) \) are bounded, it follows that Assumption B.8 holds. Assumption B.9 holds with \( H(v) = C(v) \) for some constant \( C \). Existence of four moments is assumed in Assumption RC2. Just like Assumption RC9, Assumptions B.10 and B.11 follow from Assumption RC8.

Appendix D. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2014.10.009.

References


